Universität Bielefeld

Doctoral Thesis

Optimization-Based Modeling of Suprasegmental Speech Timing

Andreas Windmann

Eigenständigkeitserklärung

Ich, Andreas Windmann, versichere, dass ich diese Dissertation,'Optimization-Based Modeling of Suprasegmental Speech Timing' selbständig verfasst habe. Ich versichere, dass

- mir die Promotionsordnung der Fakultät für Linguistik und Literaturwissenschaft der Universität Bielefeld bekannt ist;
- ich die Dissertation selbst angefertigt habe, keine Textabschnitte von Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel und Quellen in dieser Arbeit angegeben habe;
- Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Vermittlungstätigkeiten oder Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen;
- ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe;
- ich weder diese Dissertation, noch eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule als Dissertation eingereicht habe.

Datum: _____

Unterschrift: _____

Gedruckt auf alterungsbeständigem Papier ° ° ISO 9706

Mit einem Minimum an Aufwand ein Maximum an Erfolg!

Thorsten & Thomas Conrad, *Abitur für Faule*

Zusammenfassung

In dieser Arbeit wird die Hypothese überprüft, dass die suprasegmentale Zeitstruktur von gesprochener Sprache – also die Gesamtheit von Erscheinungen, die die zeitliche Ausdehnung von Silben und größeren Einheiten wie Wörtern und Phrasen in der gesprochenen Sprache betreffen – als Konsequenz eines Ökonomieprinzips verstanden werden kann: Einerseits sind Menschen bestrebt, den Aufwand bei der Sprachproduktion gering zu halten; andererseits muss für erfolgreiche Kommunikation hinreichende Verständlichkeit gewährleistet sein, was wiederum ohne Aufwand nicht zu erreichen ist. Grundannahme der Arbeit ist, dass die tatsächlich produzierten suprasegmentalen Zeitstrukturen in der gesprochenen Sprache *effizient* sind, das heißt, mit dem bestmöglichen Verhältnis von Aufwand zu Verständlichkeit produziert werden, und dass sich eine Reihe von empirisch beobachteten Timingphänomenen in der gesprochenen Sprache durch ebendiesen Umstand erklären lässt. Diese Annahme wird in der vorliegenden Arbeit in einem Computermodell formalisiert und durch Simulationsexperimente überprüft.

Die theoretischen Grundlagen für den verfolgten Modellierungsansatz werden im zweiten Kapitel der Arbeit gelegt. Das postulierte Ökonomieprinzip in der Sprachproduktion wird hier näher erläutert und es werden Befunde diskutiert, die die Plausibilität dieses Prinzips als Erklärung für eine Reihe von phonetischen und phonologischen Phänomenen untermauern. Insbesondere wird auf bereits existierende Computermodelle eingegangen, die Ökonomieprinzipien als Erklärung für verschiedene Phänomene der gesprochenen Sprache mit mathematischen Optimierungsalgorithmen formalisieren. Der Erfolg dieser Modelle untermauert das Vorhaben, den optimierungsbasierten Modellierungsansatz auf die Ebene der suprasegmentalen Zeitstruktur gesprochener Sprache auszuweiten, für die ein solches Modell bislang noch nicht existiert. In diesem Kapitel werden außerdem häufig geäußerte Kritikpunkte an ökonomiebasierten Erklärungen in der Phonetik und Phonologie diskutiert.

Das dritte Kapitel der Arbeit gibt einen Überblick über die zu modellierenden empirischen Befunde, suprasegmentale Timingphänomene in den Sprachen der Welt, wobei sich die Darstellung in dieser Arbeit größtenteils auf Betonungssprachen beschränkt. Diese Phänomene werden in vier Klassen unterteilt: Erstens Längungseffekte aufgrund von prosodischer Prominenz auf der Wort- (Betonung) und Phrasenebene (Akzent); zweitens Längungseffekte an prosodischen Grenzen; drittens Kürzungseffekte als Funktion der Anzahl der Silben in größeren prosodischen Einheiten und viertens "globale" äußere Einflüsse wie Anforderungen an das Sprechtempo oder Bedingungen, die einen besonders hörerorientierten Sprechstil erfordern. Besonderes Augenmerk wird auf die Interaktionen zwischen diesen Effekten gelegt. Ein Ergebnis des Literaturüberblicks ist,

dass die dritte Kategorie, Kürzungseffekte als Funktion der Anzahl der Silben in größeren prosodischen Einheiten, vermutlich ein Artefakt der phrasalen Prominenz ist, und damit keine eigene Kategorie darstellt. In diesem Kapitel werden auch mögliche Interpretationen der verschiedenen Kategorien mit Blick auf die postulierten Okönomieprinzipien in der Sprachproduktion vorgestellt. Die Darstellung zeigt, dass sich insbesondere Prominenzeffekte sehr gut in dieses Schema einordnen lassen, wenn man prosodische Prominenz als lokal stärkere Gewichtung der Maximierung von Wahrnehmbarkeit auffasst. Es wird argumentiert, dass Positionseffekte ebenfalls kommunikative Signale darstellen, die der Strukturierung der gesprochenen Sprache dienen; allerdings ist diese Einordnung weniger eindeutig, da in der Literatur auch die alternative Erklärung vorgebracht wurde, dass Längungseffekte an prosodischen Grenzen biomechanische Konsequenzen der menschlichen Vokaltraktphysiologie sind.

Im vierten Kapitel werden existierende Modelle diskutiert, die Erklärungen für suprasegmentale Timingphänomene anbieten. Es wird argumentiert, dass die vorgestellten Modelle in den meisten Fällen keine adäquaten Erklärungen für die beobachteten Phänomene liefern. Insbesondere gehen einige dieser Modelle davon aus, dass suprasegmentales Timing in gesprochener Sprache auf quasi-periodischen Mechanismen basiert, was mit Blick auf die Erkenntnisse aus dem dritten Kapitel als unwahrscheinlich anzusehen ist. Einige der besprochenen Modelle weisen außerdem dahingehend Unzulänglichkeiten auf, dass sie beobachtete Phänomene lediglich nachbilden, anstatt sie wirklich zu erklären.

Im fünften Kapitel der Arbeit wird die Architektur des verwendeten optimierungsbasierten Modells erläutert. Herzstück des Modells ist ein mathematischer Optimierungsalgorithmus. Dieser berechnet Silbendauern in einer simulierten Äußerung dergestalt, dass eine Kostenfunktion minimiert wird. Die Terme dieser Kostenfunktion sind ihrerseits Funktionen der Dauern von Silben und anderen prosodischen Einheiten in der simulierten Äußerung. Diese Funktionen verkörpern die hypothetischen Anforderungen an Minimierung von Aufwand und Maximierung von Verständlichkeit; außerdem lassen sich Variationen im globalen Sprechtempo als unabhängige Größe manipulieren. Gewichtungsfaktoren erlauben es, die Balance zwischen den verschiedenen Anforderungen lokal und global zu variieren und so inner- und außersprachliche Einflüsse auf die Zeitstruktur gesprochener Sprache zu simulieren. Ein zentraler Aspekt des Modelldesigns ist, dass die mathematischen Funktionen, die die Anforderungen hinsichtlich Aufwand und Verständlichkeit verkörpern, nicht willkürlich gewählt, sondern durch unabhängige Evidenz aus Sprachproduktions- und Perzeptionsforschung motiviert sind.

Bevor die Vorhersagen dieses Modells überprüft werden, werden im sechsten Kapitel die Vorhersagen einiger der im vierten Kapitel vorgestellten Modelle in den Blick genommen. Im Einzelnen werden zwei Phänomene betrachtet, die in der Literatur als Belege

für quasi-periodische Timingmechanismen herangezogen wurden. Das in dieser Arbeit verwendete optimierungsbasierte Modell sagt Nulleffekte für beide Phänomene voraus. Korpusanalysen und Simulationsexperimente, die in diesem Kapitel beschrieben werden, stützen diese Vorhersage und legen nahe, dass die betrachteten Phänomene statistische Artefakte der Verteilung bestimmter linguistischer Kategorien in der gesprochenen Sprache sind und keine periodischen Timingmechanismen als Erklärung erfordern.

Im siebten Kapitel werden zunächst empirische Analysen von gesprochenen Korpora in verschiedenen Sprechgeschwindigkeiten präsentiert. Die Analysen liefern Evidenz für die Intuition, dass Silben in der gesprochenen Sprache nicht beliebig kurz produziert, sondern nur bis zu einer bestimmten Untergrenze gekürzt werden können. Simulationsexperimente zeigen, dass das Modell dieses Ergebnis reproduziert und auch beobachtete Unterschiede hinsichtlich der Daueruntergrenzen betonter und unbetonter Silben korrekt voraussagt. Eine modifizierte Version des Modells reproduziert auch die Beobachtung, dass Silben zwar nur bis zu einer bestimmten Untergrenze gekürzt, wohl aber komplett getilgt werden können. Diese Version des Modells sagt den Einfluss der Betonung allerdings nicht korrekt voraus.

Im achten Kapitel werden schließlich Ergebnisse von Simulationsexperimenten vorgestellt, die die im dritten Kapitel beschriebenen suprasegmentalen Timingeffekte und deren Interaktionen in den Blick nehmen. Dabei zeigt sich, dass das optimierungsbasierte Modell insbesondere im Bezug auf Effekte und Interaktionen der prosodischen Prominenz sehr gute Ergebnisse liefert: Die Strategie, Prominenz als lokale stärkere Gewichtung der Wahrnehmungsseite zu modellieren, liefert korrekte Voraussagen und plausible Erklärungen für eine Reihe von prominenzbasierten Timingphänomenen. Bei der Modellierung von Positionseffekten wird ein etwas spekulativerer Ansatz verfolgt, der ebenfalls einige korrekte Voraussagen macht. Auch wenn nicht alle der im dritten Kapitel vorgestellten Ergebnisse korrekt reproduziert werden, bleibt als Gesamtergebnis festzuhalten, dass das vorgestellte Modell, besonders eingedenk seiner Einfachheit, ermutigende Ergebnisse erzielt. Das postulierte Ökonomieprinzip wird dadurch als plausibler Erklärungsmechanismus für suprasegmentale Timingmuster in der gesprochenen Sprache etabliert.

Acknowledgements

This is the optimal time and place for a few words of thanks! Throughout this work, Petra Wagner has been a great and most supportive supervisor, and I thank her wholeheartedly for being such a good mentor and friend. I am also indebted to Juraj Šimko, for inspiring this project and for highly instructive as well as entertaining collaboration. More thanks are due to my thesis committee members, Bernd Möbius, David Schlangen and Marcus Kracht, and to two further academic instructors: Jan de Ruiter, who, long ago, taught me a very insightful undergraduate course on computational models, and Dafydd Gibbon, who, even longer ago, introduced me to the world of speech science.

I gratefully acknowledge the Bielefeld Graduate School of Linguistics and Literary Studies (LiLi-Kolleg) and the graduate school of the Center of Excellence of Cognitive Interaction Technology (CITEC) for funding the research presented in this work, Dellwo et al. (2004), Auran et al. (2004) and Avanzi et al. (2010) for making their corpora publicly available, and Sunil Patel for this nice LaTeX template.

To my present and former colleagues from the Bielefeld Phonetics and Phonology group – Abir Anbari, Simon Betz, Aleksandra Ćwiek, Laura de Ruiter, Robert Eickhaus, Hendrik Hasenbein, Angelika Hönemann, Zofia Malisz, Mikhail Ordin, Leona Polyanskaya, Barbara Samlowski, Valentina Schettino, Joana Skubisz and Marcin Włodarczak: thanks for the great time!

Ich danke meinen Eltern, die immer für mich da waren. Außerdem danke ich meinen Freunden in der LKG Bielefeld-Mitte und in der SMD Bielefeld, wo ich in den vergangenen Jahren viele fröhliche Stunden verbringen und eine geistliche Heimat finden durfte.

Ich vermag alles durch den, der mich stark macht, Christus. (Phil. 4,13)

Contents

Eigenständigkeitserklärung **i**

Zusammenfassung **iii**

Acknowledgements **vi**

Contents **vii**

List of Figures **x**

List of Tables **xii**

1 Introduction **1**

I Theoretical Background **7**

2 Efficiency-Based Explanations of Speech Patterns **8**
2.1 Introduction . . . 8
2.2 Review . . . 9
2.2.1 Hyper- and Hypoarticulation Theory . . . 9
2.2.2 Optimality Theory and Related Approaches . . . 16
2.2.3 Embodied Task Dynamics . . . 26
2.2.4 Other Approaches . . . 33
2.2.5 Criticisms of Efficiency-Based Explanations in Speech Science . . . 37
2.3 Discussion . . . 39

3 Suprasegmental Speech Timing **41**
3.1 Introduction . . . 41
3.2 Prominence Effects . . . 42
3.2.1 Introduction . . . 42
3.2.2 Review . . . 44
3.2.3 Summary . . . 49
3.3 Positional Effects . . . 50
3.3.1 Introduction . . . 50
3.3.2 Review . . . 51
3.3.3 Summary . . . 55

3.4 Constituent Length Effects . . . 56
3.4.1 Introduction . . . 56
3.4.2 Review . . . 56
3.4.3 Summary . . . 61
3.5 Effects of Overall Speaking Rate . . . 62
3.5.1 Introduction . . . 62
3.5.2 Review . . . 62
3.5.3 Summary . . . 67
3.6 Overall Summary . . . 68

4 Explanatory Accounts of Suprasegmental Speech Timing 70
4.1 Introduction . . . 70
4.2 Review . . . 70
4.2.1 Oscillatory Models . . . 70
4.2.2 The Converter/Distributor Model . . . 82
4.2.3 Other Approaches . . . 84
4.3 Discussion . . . 89

II Model Definition and Results 90

5 Model Definition 91
5.1 Introduction . . . 91
5.2 Model Components . . . 93
5.2.1 Effort Cost E . . . 93
5.2.2 Perception Cost P . . . 97
5.2.3 Duration Cost D . . . 103
5.3 Optimization . . . 104
5.4 Discussion . . . 107

6 Testing Predictions of Models of Speech Timing 110
6.1 Introduction . . . 110
6.2 A study of Constituent Length Effects in English . . . 111
6.2.1 Introduction . . . 111
6.2.2 Corpus Analysis . . . 112
6.2.2.1 Material and Methods . . . 112
6.2.2.2 Results . . . 114
6.2.3 Discussion . . . 118
6.3 Investigating Regression Results on Inter-Stress Interval Duration . . . 122
6.3.1 Introduction . . . 122
6.3.2 Analysis . . . 123
6.3.3 Discussion . . . 130
6.4 General Discussion . . . 131

7 Incompressibility 132
7.1 Introduction . . . 132
7.2 Corpus analyses . . . 134
7.2.1 Data and Method . . . 134

7.2.2 Results . . . 135
7.2.3 Discussion . . . 139
7.3 Modeling Durational Incompressibility . . . 141
7.3.1 Basic Model . . . 141
7.3.2 Modified Model . . . 144
7.4 Discussion . . . 150

8 Modeling Effects of Prominence, Position and External Conditions on Suprasegmental Speech Timing 152
8.1 Introduction . . . 152
8.2 Prominence Effects . . . 153
8.2.1 Interaction of Stress and Accent . . . 153
8.2.2 Constituent Length Effect in Accented Words . . . 156
8.3 Effects of Overall Speaking Rate: Time Constraints and Global Hyperarticulation . . . 160
8.4 Positional Effects . . . 162
8.5 Discussion . . . 167

9 Conclusion 169

Bibliography 174

Appendix A: Source Code of the Optimization-Based Model of Speech Timing 194

Appendix B: Source Code of the Mass-Spring Model 198

List of Figures

2.1 Target undershoot model. 11
2.2 Schematic overview of ETD. 27
2.3 Plot of the temporal realization estimate in ETD. 29
2.4 Plot of /abi/ and /iba/ simulations in ETD. 31
2.5 Articulatory and voicing effort in the Elija model. 35

4.1 Syllable triangles in the C/D model. 83

5.1 Articulatory effort estimate. 95
5.2 Articulatory, phonatory and overall effort in the model. 97
5.3 Cost function E. 98
5.4 Plot of the temporal perception cost in ETD. 99
5.5 Recognition scores from gating studies. 100
5.6 Plot of perceptual cost function P. 102
5.7 Evolution of overall cost of simulated utterance during optimization. . . . 106
5.8 Syllable durations predicted by the model. 107
5.9 Model architecture. 108

6.1 Vowel duration by prominence and within-word position in the Aix-MARSEC corpus. 114
6.2 Vowel duration by syllable count in the ISI in the Aix-MARSEC corpus. . 115
6.3 Vowel duration by syllable count in the word in the Aix-MARSEC corpus. 116
6.4 vowel duration by syllable count in the NRU in the Aix-MARSEC corpus. 117
6.5 Percentage of word-final observations as a function of constituent length in the Aix-MARSEC corpus. 118
6.6 Simulation of MARSEC corpus data. 121
6.7 Percentage of stressed syllables that are word-final as a function of syllable count in the ISI in English, French and Finnish. 125
6.8 Regression results on simulated English, French and Finnish data. 127
6.9 ISI duration by number of syllables in Finnish, from O'Dell and Nieminen (2001). 128
6.10 Simulated regression intercepts and stressed-unstressed duration differences in English, French and Finnish. 129

7.1 Illustation of Klatt (1973)'s duration model. 133
7.2 Fast∼normal regression on BTC data. 136
7.3 Fastest possible∼normal regression on BTC data. 137
7.4 Fast∼normal regression on PC data. 140
7.5 Fast∼slow regression on simulated data. 142

7.6 Illustrations of the incompressibility effect in the model. 143
7.7 Fastest possible∼normal regression on English unstressed BTC data. . . . 145
7.8 Plot of modified perceptual cost function P. 145
7.9 Fast∼slow regression on simulated data using the modified model. 146
7.10 Plot of modified cost function C. 147
7.11 Fast∼slow regressions on simulated data for increasingly fast rates with stress distinction. 149
7.12 Plot of cost function C for stressed and unstressed syllable. 150

8.1 Syllable durations predicted by the model. 153
8.2 Accentual lengthening predicted by the model. 154
8.3 Polysyllabic shortening in accented contexts as predicted by the model. . 158
8.4 Total accentual lengthening as a function of word length as predicted by the model. 159
8.5 Simulation of speaking rate effects. 160
8.6 Simulation of H&H scale variation effects. 163
8.7 Simulation of final lengthening in stressed and unstressed syllables. 164
8.8 Simulation of final lengthening in accented and unaccented stressed monosyllables. 165
8.9 Simulation of word-final lengthening in accented words. 166
8.10 Simulation of final lengthening under speaking rate variation. 167

List of Tables

6.1 Number of syllables by syllable count in the ISI in the Aix-MARSEC corpus (English) and the C-PROM corpus (French). 123
6.2 Regression models of ISI duration by number of component syllables in English, French and Finnish. 124
6.3 Assignment of simulated stressed non-final and stressed final syllables by ISI length in English, French and Finnish. 126

7.1 Syllables and speakers per language and actual speaking rate by tempo condition in the BonnTempo Corpus. 135
7.2 Summary of speaking rate regression models of stressed and unstressed syllable duration from the BTC. 135
7.3 Regression summary of fast rate on slow rate durations by syllable type in the Petra Corpus. 139

Chapter 1

Introduction

This thesis is concerned with the optimal use of time in talking. It investigates the hypothesis that patterns of suprasegmental speech timing – i.e., the timing of syllables and larger units in speech – emerge from trade-offs between the conflicting demands of minimizing production effort and maximizing perceptual clarity, or, ultimately, communicative success. The idea that such principles are a governing factor in human speech production is not new; it has been formulated most prominently in the Hyper- and Hypoarticulation (H&H) theory by Lindblom (1990). As we shall see, a large body of evidence suggests that similar principles provide promising explanations for a wide range of phenomena in human speech production and other domains. Our contribution will be to show that this also applies to suprasegmental speech timing, for which currently no such unified account exists.

As such, this work could be conceived as a purely theoretical enterprise, involving a review of empirical results and theoretical speculations about their interpretation. However, this methodology is bound to reach its limits when it comes to deriving actual predictions: for example, one team of researchers cited in this work argues that H&H theory predicts stressed vowels to shorten less strongly under increased speaking rate than unstressed vowels, while we argue the exact opposite. Both hypotheses can be well motivated by theoretical reasoning and are thus equally valid *a priori*. External evidence is necessary to decide between them. For this reason, we have chosen to adopt a different strategy: the hypothetical goals to minimize effort and maximize communicative success in speech production have been formalized in a computational model. Predictions made by this model will be evaluated against attested speech timing patterns. The adequacy of our initial hypothesis can then be established by the capacity of the model to reproduce empirically observed phenomena.

How can our assumptions be implemented? We think it reasonable to assume that any given speech event cannot perfectly satisfy both demands at the same time: the least effortful behavior, for example, is obviously to do nothing and remain silent, but this is also the worst choice for communicative success. However, it may be possible to produce speech such that it is *efficient*, meaning that it exhibits the highest possible "clarity-to-effort ratio", given that clarity and effort are parametrized in a meaningful way. Nonlinear optimization algorithms present a powerful method for formalizing this problem. Two prerequisites are necessary for this, 1) a device that simulates speech production, or at least generates representations of speech events which embody the aspect of speech that is of interest, and, 2) as stated above, well-defined metrics that score simulated speech events with regard to how well they satisfy the hypothetical demands efficiency, i.e., minimizing effort while maximizing perceptual clarity.

Once these requirements are in place, the degree of satisfaction of the demands for minimizing effort and maximizing clarity can be implemented as mathematical cost functions, whose particular design should be based on independently motivated assumptions about speech production and perception. The optimization algorithm then determines the parameters of the production model such that the simulated speech output incurs the lowest overall cost, corresponding to the optimal trade-off between minimizing effort and maximizing clarity, and, hence, a maximally efficient speech output. Numerical weights can be used in this approach to simulate global or local conditions that modulate the relative importance of effort minimization and clarity maximization. If our assumptions are correct, then this speech output form should exhibit characteristics that are also observed in real speech.

Are our assumptions realistic? Intuitively, the idea that speakers strive for successful communication would seem to be a truism. Likewise, to assume that humans should possess a predilection for minimizing effort appears plausible, "as displayed by the popularity of chairs and automobiles", as (Kochanski and Shih 2003:324) put it. Moreover, the plausibility of our assumptions is supported by evidence from non-speech domains, which shows that similar explanations based on trade-offs between minimization of metabolic energy expenditure and the necessity to perform certain actions account for a range of motor and behavior patterns in living organisms. A famous example has been provided by Hoyt and Taylor (1981): the authors put horses on a treadmill and measured oxygen consumption per traveled distance at different gaits. In subsequent observations, they found that horses running freely in a paddock tended to move at precisely those speeds that had been found to minimize oxygen consumption at the respective gaits in the treadmill experiment. Thus, it appears that locomotion in horses is optimized for metabolic energy expenditure as measured by oxygen consumption. In wildlife biology, *Optimal Foraging Theory*, which claims that animals attempt at minimizing gain and

maximizing costs such as time consumption in searching for food, has been demonstrated to account for a range of observations. (see Smith 1982 for an overview). Modeling work by Anderson and Pandy (2001) has shown that similar principles may apply to locomotion in humans, as demonstrated by the reproduction of several physiological characteristics of human walking in an optimization model that derives solutions for motion equations by minimizing different measures of energy expenditure. Further examples of such studies are listed in the review by Todorov (2004). More importantly, we will see that efficiency-based approaches have been successful in accounting for various speech-related phenomena. At the same time, there are also researchers who have disputed the adequacy of such accounts. An integral part of our argument, therefore, will be to assess evidence for efficiency-based explanations in speech science, as well as potential counterarguments.

Before introducing the structure of this thesis, some fundamental methodological reflections are in order. Although it is concerned with similar phenomena, the computational model of speech timing we are going to introduce here belongs to an entirely different class of models than the duration models commonly employed in speech technology applications. The crucial difference is that most of these models are purely *descriptive*. The task of such models is, unsurprisingly, to provide the best possible description of durational data. They do so by directly approximating observed durations of speech events, utilizing any suitable mathematical technique. The goodness of these models is measured by assessing how well they perform in minimizing the numerical error between observed and predicted timing patterns. To date, various statistical techniques, such as decision trees, neural networks or regression models are available for predicting speech timing based on linguistic and phonetic features. Importantly, no theoretical status is attached to the components of descriptive models, and the decision whether to include a particular feature is guided only by its potential to improve prediction.

While such descriptive models are often very successful at predicting speech timing patterns, they "do not embody fundamental insights in the communicative, linguistic, physiological and acoustic processes underlying temporal patterning in speech" (Nooteboom 1991:230). For example, a linear regression model may yield a very accurate description of the commonly observed positive correlation between vowel duration and the degree of jaw opening, but it does not provide an account of the underlying physiological reasons for this relationship. This is of course entirely unproblematic in the applications where descriptive models are used, because underlying mechanisms are usually not of interest in these applications as long as surface observations are faithfully reproduced. An explanatory account, on the other hand, requires hypotheses about precisely such underlying mechanisms. As for the relationship between jaw opening and vowel duration, the work by Lindblom (1967) represents an attempt of this kind: in this approach,

the jaw is modeled as a mass-spring system, representing different degrees of opening by the displacement of the mass. The *explanation* this model yields for the effect of jaw opening on duration is that with a larger displacement, it takes longer for the spring to return to its resting position, provided that the parameters of the spring are kept constant.

Thus, to recapitulate, while descriptive models approximate observed phenomena directly, explanatory models approximate hypothesized underlying mechanisms, with the goal that empirically observed phenomena emerge automatically from the implementation of these mechanisms. If this happens, the model can be said to provide a sufficient (although by no means a necessary) *explanation* of the phenomenon under study. Explanatory models are thus also concerned with minimizing a prediction error of sorts, in the sense that the model output should "resemble" the empirical observation along some dimension for the model to be judged adequate. Yet, this resemblance is typically evaluated at a much coarser level for explanatory than for descriptive models; the focus of explanatory models usually lies on reproducing *qualitative patterns*, rather than on matching exact numerical results. For example, in the (Lindblom 1967) model, the crucial outcome is the general relationship between displacement and oscillation period in a mass-spring system, which replicates the relationship between jaw displacement and vowel duration. No further insights would be gained from fitting spring parameters to match any particular set of durational measurements, not least because it is not clear *what* particular set of measurements should serve as reference for the model.[1]

The boundary between descriptive and explanatory models is not always clear cut. First, despite the above considerations, it is of course possible to design models that achieve both descriptive and explanatory adequacy. For example, the intonation models by Fujisaki and Hirose (1984) and Prom-On et al. (2009) are based on hypothesized physiological production mechanisms and yet can be used to generate numerical predictions of intonation contours in actual applications. Prom-On et al. (2009) give an impressive demonstrations of this by showing how the phenomenon of *peak delay* – the property of intonation contours to reach their peak *after* the accented syllable – emerges automatically from the physiologically motivated modeling assumptions, without the need to prescribe it by a dedicated parameter. Second, some models that purport to be explanatory are actually rather descriptive. The philosophical problem that is at issue here is circularity: if the very results that the model reproduces are used to motivate modeling assumptions, then very little is achieved – the model in this case merely demonstrates that a phenomenon can be *implemented* using a particular modeling technique,

[1] This is not to say that an explanatory model would not be judged better than another one that matches observed data less well. The crucial point is that the better fit must be achieved by independently motivated model components, not by tuning parameters to match observations.

but it does not provide a satisfactory explanation (Turk and Shattuck-Hufnagel 2014b). Explanatory models, in other words, "must necessarily invoke information (explanans principles) independent of the facts observed (the explananda) to avoid circularity and to count as genuine explanations" (Lindblom and Engstrand 1989:109). In our model, the hypothesized tendencies towards minimizing effort and maximizing perceptual clarity represent this independent information. An important point of our argument will be to assess the plausibility of our own modeling assumptions, as well as those that underlie other explanatory models of speech timing.

The remainder of this thesis is structured as follows: Part I covers the theoretical background. It starts with a review of efficiency-based accounts of speech patterns in Chapter 2. We shall provide a thorough discussion of H&H theory and discuss theoretical, experimental and model-based accounts that are built upon similar principles. In Chapter 3, we present a concise overview of the empirical phenomena to be modeled, suprasegmental timing patterns in speech. The theoretical exposition is completed by Chapter 4, where we review and critique existing explanatory accounts of suprasegmental speech timing. In Part II of this thesis, we introduce the design of our model and present results of simulation experiments and empirical studies on corpus data that explore its predictions. It starts out with Chapter 5, which contains the formal definition of our model. The individual model components are introduced and motivated, and the optimization procedure is described. Before examining predictions made by our model, we focus on some of the alternative models introduced in our review: two phenomena predicted by these models, for which our model predicts null results, are investigated using corpus analyses and simulation experiments in Chapter 6. We show that both effects are likely to be artifacts of language structure, and do not require dedicated timing mechanisms. Chapters 7 and 8, finally, deal with predictions of our own model. In Chapter 7, we investigate the low-level phenomenon of durational incompressibility in speech on corpus data and report simulation experiments showing that our model provides a convincing account of the observed durational patterns. In Chapter 8, we return to timing phenomena arising from higher-level linguistic structure as reviewed in Chapter 3 and demonstrate their reproduction in our model. A general discussion of our findings and some concluding remarks are provided in Chapter 9.

Parts of this thesis and the work presented therein have been published in the following articles:

- Windmann, A., Šimko, J., Wrede, B., & Wagner, P. (2013). Modeling durational incompressibility. *Proceedings of Interspeech 2013*, Lyon, 1375–1379.
- Windmann, A., Šimko, J., & Wagner, P. (2014). Probing theories of speech timing using optimization modeling. *Proceedings of Speech Prosody 7*, Dublin, 346–350.

- Windmann, A., Šimko, J., & Wagner, P. (2014). A unified account of prominence effects in an optimization-based model of speech timing. *Proceedings of Interspeech 2014*, Singapore, 159–166.

- Windmann, A., Šimko, J., & Wagner, P. (2015). What do regression analyses of inter-stress interval duration really measure? *Proceedings of ICPhS 2015*, Glasgow, A-66.

- Windmann, A., Šimko, J., & Wagner, P. (2015). Polysyllabic shortening and word-final lengthening in English. *Proceedings of Interspeech 2015*, Dresden, 36–40.

- Windmann, A., Šimko, J., & Wagner, P. (2015). Optimization-based modeling of speech timing. *Speech Communication* 74, 76–92.

Part I

Theoretical Background

Chapter 2

Efficiency-Based Explanations of Speech Patterns

2.1 Introduction

In this chapter, we shall review evidence pertaining to efficiency-based explanations and models of speech patterns. The term "efficiency" here is used as a shorthand for our assumption that speech patterns are shaped by the resolution of trade-offs between minimizing production effort and maximizing perceptual clarity (an alternative formulation would be to speak of *economy* principles, and we will use both terms interchangeably throughout this work). We will provide a concise and somewhat selective overview of theoretical, empirical and computational approaches in speech science that are based on similar principles, paying special attention to existing implementations of optimization-based models. We will start by reviewing Lindblom (1990)'s Hyper- and Hypoarticulation theory, which represents the most prominent formulation of efficiency principles in speech. Then, we will proceed to introduce approaches formulated within the theoretical frameworks of Optimality Theory and Articulatory Phonology, and discuss some works we failed to find a more specific common label for but still consider relevant for our argument. A separate subsection will be devoted to criticisms of efficiency-based explanations of speech patterns. The goal of the chapter is to show how the notions of minimizing production effort and maximizing perceptual clarity can be conceptualized, and, ultimately, to provide evidence from a variety of subfields of speech research in order to show that these assumptions provide promising explanations for a variety of speech-related phenomena.

While we regard Lindblom (1990) as the seminal theoretical contribution in terms of efficiency-related accounts of speech patterns, similar ideas have been discussed earlier.

Kul (2007) provides a useful historical overview of related work. One particularly prominent example mentioned in her review is the work by Zipf (1935, 1949). Zipf states that a *principle of least effort* underlies all human behavior (indeed, he claims it to be a basic operating principle of the physical universe), and, in the 1935 work, argues this point specifically for language, using evidence such as phoneme frequency patterns in support of his proposition. The hypothesis advanced in this work obviously points in a similar direction, but we would claim that "least effort" is only half the picture and must be considered in relation to the other side of the equation, maximizing communicative success. It is clear from the elaborations of Zipf and other authors who propose similar ideas that consideration of communicative success is to some extent implied by their thinking – as we said above, the least effortful speech-related behavior is, obviously, to remain silent, whereas Zipf's principle of least effort probably refers to something like "the least effort that still allows for communication to succeed". In any case, we deem it important that both aspects are explicitly addressed, and in what follows, we will concentrate on approaches that take both effort- and clarity-related criteria into account.

2.2 Review

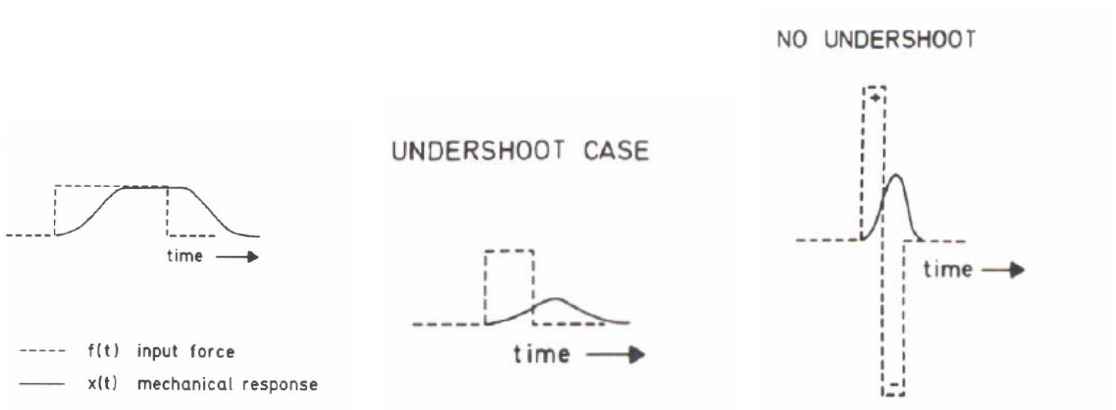
2.2.1 Hyper- and Hypoarticulation Theory

Hyper- and Hypoarticulation (H&H) theory as formulated by Lindblom (1990) provides the central theoretical foundation of our work. Its essence can be briefly summarized: human speech production, according to this theory, is characterized by "a continual tug-of-war between demands on the output on the one hand and system-based constraints on the other" (Lindblom 1990:420): on the one hand, "unconstrained, a motor system tends to default to a low-cost form of behavior' referred to as *hypoarticulation* in the speech domain (Lindblom 1990:413). On the other hand, successful transmission of the information conveyed by a given utterance has to be ensured, which requires orientation towards the listener's demands. Lindblom (1990) uses the term *hyperarticulation* to refer to a clear and listener-oriented speaking style, which, by hypothesis, requires effort on part of the speaker. The core tenet of H&H theory is that humans produce speech such as to optimally satisfy these conflicting requirements. The variation in speech, on this view, arises because different linguistic and extra-linguistic conditions under which speech is produced modulate the relative importance of both requirements. These conditions may apply at any level of speech description, from entire utterances or discourses down to, possibly, individual gestures.

Lindblom (1990) exemplifies some global production- and perception-related conditions that favor hypo- or, conversely, hyperarticulation. On the speech production side, he mentions physiological and cognitive factors. One may hypothesize that for example speaker age, health, physical exhaustion or situational cognitive load influence speech production by requiring the conservation of effort to a greater or lesser degree. As for the perceptual dimension, Lindblom names social and communicative variables such as "channel, listener, situation, degree of formality" (1990:419) as relevant variables for introducing variation on the continuum between hyper- and hypoarticulation. For example, it may be assumed that talking to an unfamiliar or non-proficient listener or in a formal setting will prompt speakers to lean more strongly towards hyperarticulation.

Evidence for the assertion that speakers default to low-cost behavior comes from an older result by the same author: Lindblom (1963) measured formant frequencies of vowels in uniform consonantal contexts across a range of speaking rates, which were elicited by asking a speaker to temporally align productions with a periodic auditory stimulus. Lindblom (1963) found that hypothetical formant targets for the vowels were systematically undershot under time pressure, indicating that the relevant articulators in these cases did not fully reach their targets. Lindblom (1990) also discusses articulatory evidence from a study by Nelson et al. (1984): these authors had subjects produce alternating jaw opening and closing movements as well as repetitive /sa/ syllables at increasing rates. Nelson et al. (1984) found movement amplitudes to decrease with increasing rate. Lindblom (1963) and Nelson et al. (1984) put forward a common explanation for their findings: speakers avoid very high movement velocities, as would be required to reach articulatory targets in short time intervals. Nelson et al. (1984) explicitly hypothesize that the peak velocity of articulatory movements can be used as a measure of physical effort, and conclude that speakers tend to produce speech in such a way that effort is conserved.

Further studies reviewed by Lindblom (1990), however, have put these results into perspective. He refers to findings (e.g. Gay 1978, Kuehn and Moll 1976) showing that *duration-dependent undershoot* as observed by Lindblom (1963) does not necessarily occur; speakers are capable of fully reaching articulatory targets in very short time. H&H theory would explain this result as arising from the simultaneous necessity to ensure successful communication. Thus, speakers can willfully invest more effort if speaking conditions necessitate it. Lindblom (1983) illustrates this *plasticity* property of the speech production system using a mechanical model of a single articulator, in which a force is applied to a mass attached to a damped spring. Such a system "moves sluggishly in response to a force that is applied and removed abruptly" (1983:227), as can be seen in panel (a) of Figure 2.1. Due to this sluggishness (which is a consequence of the damping), the response of the system will fail to attain the movement target if the



(a) (b) (c)

FIGURE 2.1: Target undershoot in a mass-spring model of articulation (adapted from Lindblom (1983). Shown are input forces and system responses. Panel (a): Input force acting upon the system (solid line) and (damped) system response (dashed line). Panel (b): undershoot due to short time interval. Panel (c): Avoidance of undershoot by increasing input force (and introducing opposite impulse).

time interval during which the force is applied is very short, as shown in panel (b) of Figure 2.1. This target undershoot can be avoided by increasing the force impulse and adding an opposite impulse "to bring the system back on time" (Lindblom 1983:230), as shown in panel (c) of Figure 2.1. Thus, target undershoot can be prevented by spending extra effort, if perceptual constraints require it. (Lindblom 1983:231) states that "(t)he system is indeed capable of raising the level of its performance, but as any phonetician will testify, it "prefers" not to".

A case in point for this assertion, according to Lindblom (1983), are coarticulation phenomena in speech, exemplified by Öhman (1966)'s data on formant transitions in VCV-sequences. These data suggest that in normal speech, consonant articulations are displaced towards vowel targets. Lindblom (1983, 1990) interprets this pattern as arising from *synergy constraints* between the tongue body and the tongue tip that facilitate the avoidance of large displacements, and, hence, secure the achievement of motor economy. However, results from bite-block experiments (Lindblom et al. 1979) show that even with an object in the mouth that enforces an unusually large jaw opening, speakers are capable of producing intelligible renditions of the required utterances by using extraordinarily large, compensatory tongue movements. (Lindblom 1983:224) states that "(n)ormal speech seems to exploit no more than a fraction of the degrees of freedom that are in principle available for articulation". If necessary, however – as in the case of a bite block experiment – the system is capable of instantiating such large displacements, to ensure that perceptual requirements are met.

Lindblom (1990) discusses more evidence to demonstrate how H&H variation caused by changes in external conditions influences the speech signal. Some particularly intriguing examples com from singing: Lindblom (1990) cites a study by Johnson et al. (1983), which shows that singers use a tongue configuration resembling that of an /a/ when producing /u/ and /i/ vowels when the fundamental frequency (F0) exceeds the first formant. The interpretation in terms of H&H theory is that with an F0 this high, perceptual cues for F1 are unavailable anyway, and singers can revert to the tongue configuration for /a/, which Lindblom deems the most neutral of the three vowels. Another interesting finding concerns the "singer's formant", a term which denotes the proximity of the third, fourth and fifth formant typically found in voiced sounds sung by classical singers. Lindblom (1990) cites research by Sundberg (1987), who argues that this is a feature deliberately used by singers to increase the perceptual contrast between their singing and the accompanying orchestra, whose spectrum does usually not include a peak in this region. The singer's formant is, however, not found in soprano singing, which Sundberg (1987) attributes to the fact that the spectrum of soprano singers is by default more distinct from that of the orchestra due to their higher pitch range. Lindblom (1990) interprets this as an instance of H&H variation: there is no need for soprano singers to instantiate the vocal tract modifications necessary to produce the singer's formant, because their singing is distinctive enough without it.

A crucial assumption of H&H theory that is related to output-oriented considerations is that speech perception is aided by *signal-complementary processes*. Lindblom (1990) discusses results by Luce (1986) as an example of this. Luce (1986)'s *Neighborhood Activation Model* predicts spoken word recognition to be a function of both the frequency of the word itself as well as the density of the neighborhood, i.e., the number of acoustically similar words in the lexicon. Luce (1986)'s results cited in Lindblom (1990) suggest that these predictions are borne out by experimental data. H&H theory assumes that speakers are aware of such additional sources of information that are available to listeners, and adapt their productions to these circumstances. Thus, H&H theory would predict highly frequent words with low neighborhood densities to be particularly prone to acoustic reduction, because for these words, there is a low probability of confusion, and speakers can accordingly afford to reduce them. We will discuss studies that provide support for similar predictions later in this chapter.

Claims about the importance of signal-complementary processes relate H&H theory to one of the most fundamental problems in speech research, the *invariance problem*. The essence of this problem is the question how listeners can extract linguistic categories such as phonemes or words from the speech signal despite the massive contextual variation in speech, or, in other words, the apparent lack of invariant acoustic correlates of linguistic

categories. Various solutions to this problem have been proposed and discussed in Lindblom (1990), some of which claim that invariance resides in the articulatory domain, which is directly perceived by listeners (*Motor Theory*, Liberman and Mattingly 1985; *Direct Realism*, Fowler 1986), while others assert that invariance is indeed to be found in the acoustic domain (Miller 1989, Stevens 1986). H&H theory takes an entirely different stance towards this problem (Lindblom 1990:431):

> In contrast, the H&H theory assumes that, in all instances, speech perception is the product of both signal-driven and signal-independent information, that the contribution made by the signal-independent processes show short-term fluctuations, and that speakers adapt to those fluctuations. It says that – whether communicatively successful or not – that [sic] adaptive behavior is the reason for the alleged lack of invariance in the speech signal. Hence it predicts that the quest for signal-based definitions of invariance will continue to remain unsuccessful as a matter of principle. In the H&H model the need to solve the invariance issue disappears. But the problem is replaced by another [...]: That of describing the class of speech signals that satisfy the condition of "sufficient discriminative power".

A core tenet of H&H theory is that regularities commonly ascribed to phonological rules originate from motor economy principles. Lindblom exemplifies this for the frequent class of phonological assimilations in the languages of the world: "(a)n assimilation [...] invariably implies shortened movement (glottal or subglottal). If once more we [...] examine the efficiency of such a system in terms of energy expenditure, we see that assimilation, defined as reduced distance between two sequentially timed articulatory targets, implies less work per unit time" (1983:237). Lindblom (1983) furthermore posits that properties of the syllable as an organizational unit in speech may emerge from trade-offs between effort minimization and maximization of perceptual clarity. He discusses an intriguing interpretation of the *sonority hierarchy*, which describes the fact that within a syllable, "consonants in clusters vary with respect to their preferred distance to the sonority peak (vowel)" (1983:240 f.). Lindblom reports results from an articulatory study showing that the degree of jaw opening during consonant articulation in VCV sequences correlates with their sonority, i.e., distance to the vowel. Lindblom takes this finding to suggest that syllable structure has evolved to support the efficient production of speech by facilitating coarticulation: "*Segments that are more difficult to coarticulate show up in positions remote from each other, whereas more compatible sounds tend to be relatively more adjacent in the syllable*"(Lindblom 1983:241; italics in original).

Lindblom (1983) and Lindblom (1990) also discuss evidence from the typology of sound systems. For instance, Lindblom (1983) notes that the languages of the world seem to

prefer non-dorsal articulations for consonants, whereas vowels are mostly dorsal. His interpretation of this patterning is that "the preferred consonant-vowels (CV) sequence is one that makes temporal overlap of adjacent gestures possible" (1983:240). Lindblom (1983) speculates that simultaneously, the alternation of acoustic segments with high-pass (consonants) and low-pass (vowels) characteristics may be perceptually beneficial. Lindblom (1990) also discusses an observation on cross-linguistic tendencies in phoneme inventories: the incidence of *elaborated* and *complex* phoneme segments, i.e., segments with one and two superimposed secondary articulations, respectively, correlates with phoneme inventory size. Lindblom links this observation to economy principles: in small phoneme inventories, basic articulations are sufficient for discrimination, and more complex articulations can be avoided. In larger inventories, the perceptual space is more "crowded", hence secondary and tertiary articulations have to be invoked in order to ensure sufficient perceptual contrast.

Finally, earlier work by Lindblom raises the possibility that, beyond economy or optimization principles as a mechanism for selecting among phonological categories, the categories themselves may be a product of optimization. This idea is developed in Liljencrants and Lindblom (1972), which to our knowledge represents the earliest example of computationally implemented optimization modeling in linguistics. Liljencrants and Lindblom (1972) hypothesize that maximization of perceptual contrast is a universal organizing principle in vowel systems and set up a numerical model to test this hypothesis: they define a mel-scaled acoustic space delimited by the first two formant frequencies that corresponds to the acoustic space typically found in human speech and implement an algorithm that, for a given number of vowel categories, determines their first (F1) and second (F2) formant values such that the sum of the squared acoustic distances between all pairs of vowels is maximized. The algorithm is initialized by placing the vowels at arbitrary equidistant points in the acoustic space. (Liljencrants and Lindblom 1972:842) give the following informal explanation of the optimization algorithm, which we quote in full here because we will encounter similar algorithms in various places in this work:

> First the point is moved a certain distance, and checked for being still inside the boundary; if this is the case, a new value of E [the reciprocal of the summed acoustic distances, A.W.] is computed. This is repeated for a number of directions, usually six, out from the original location of the point. Then the optimum direction is selected, and the point is repeatedly adjusted in this direction until either the boundary is hit or E no longer decreases. Then another direction search is made, and so on until a minimum for this point is established. This whole procedure is repeated for all the points.

> After this has been done once, it is started from the beginning one or more times until E does not decrease any more.

Liljencrants and Lindblom (1972) assign broad phonemic transcriptions to the F1F2 pairs predicted by their model and compare them to preliminary survey data on vowel systems in the languages of the world. They report that the model delivers fairly accurate predictions of actually occurring inventories with a given number of vowels, especially for small inventories. For example, the model, quite unsurprisingly, predicts the three most peripheral vowel qualities, transcribable as /a/, /i/ and /u/, for inventories with three vowels, and this is also a frequently occurring configuration found among languages that distinguish between three phonemic vowel qualities. For larger inventories, observed systems are matched less well in some cases, and the comparison to the survey data is somewhat speculative in that these consist of purely impressionistic transcriptions rather than acoustic measurements, but overall, the predictions of the model match the empirical data reasonably well.

Lindblom (1986) presents some refinements to the initial model, introducing more realistic assumptions about auditory processing in humans. He shows that with these modifications, the model generates somewhat more realistic predictions. Lindblom (1986) concedes that correspondence between the model's predictions and attested inventories is still far from perfect, and he also acknowledges that many potentially relevant factors other than contrast in the F1F2 space are not taken into account. Moreover, in an ideal model, the size of the inventory would also be an emergent property, rather than a fixed parameter. Finally, Liljencrants and Lindblom (1972) mention the problem that the optimization algorithm is not guaranteed to produce a *globally* optimal solution; the solution found by the algorithm "merely represents that system which exhibits greater over-all contrast then the other systems examined during the computations" (1972:855). This is indeed a general problem of optimization algorithms. An obvious question, finally, is why some languages have vowel systems that are apparently non-optimal. We will discuss related questions later in this chapter.

Despite these reservations, the modeling approach by Liljencrants and Lindblom (1972) and Lindblom (1986) provide an intriguing demonstration of the explanatory potential of H&H assumptions. In particular, these works show how linguistic categories may emerge from more general properties of human communication. Thus, they place H&H theory in a larger context of *substance-based* theories of linguistics, and in contrast to formal approaches of the generative tradition (Chomsky 1957) which adhere to the view that linguistic forms are arbitrary and explicitly deny that they could originate from some more fundamental biological, psychological or communicative principles. In our opinion, a substance-based account as advocated by H&H theory represents a more

fruitful approach to the study of language: as we said in the introduction to this work, only by invoking more fundamental principles that are rooted in human biology and psychology will it be possible to derive genuine *explanations* of observed phenomena. In the remainder of this chapter, we will review further evidence for the plausibility of H&H assumptions in speech.

2.2.2 Optimality Theory and Related Approaches

Optimality Theory (OT; Prince and Smolensky 2008) may be viewed as a generalization of the principles underlying H&H theory: it utilizes the same mechanism, namely the idea that linguistic forms are shaped by the interaction between conflicting constraints, but is more general in that the constraints are not necessarily related to effort minimization and maximization of perceptual clarity. Instead, OT posits that *faithfulness* and *markedness* constraints interact in the shaping of linguistic structure. The markedness of a linguistic entity relates to its being "more complex than an alternative along some dimension" (Prince and Smolensky 1997:1605), whereas the faithfulness criterion requires that a linguistic entity to be produced be as similar as possible to its hypothesized underlying representation. In short, OT assumes that linguistic surface forms are determined by evaluating all possible forms against the relevant constraints. It is assumed that these constraints are hierarchically ranked. The candidate that violates the lowest-ranking constraints – i.e., that is *optimal* with regard to the set of constraints – is the winner and surfaces in the speech output.

Faithfulness and markedness do not correspond directly to hyper- and hypoarticulation, but it is obvious that the general mechanisms of both theories are compatible, and some proponents of OT have employed constraints that are directly motivated by minimization of effort and maximization of perceptual clarity as functional principles in speech production. A prominent approach of this kind, Flemming (2001b), focuses on the role of auditory representations in phonology. His "dispersion theory of contrast", in line with H&H theory, claims that sound inventories are shaped by three principles, (1) maximization of the number of contrasts, (2) maximization of the distinctiveness of contrasts, and (3) minimizing articulatory effort. The first two principles can be related to hyperarticulation or at least to output-oriented control, whereas the third principle is identical to Lindblom's hypoarticulation. Flemming (2001b) utilizes interaction of constraints based on principles (1)–(3) in phonological analyses of a variety of sound patterns from the languages of the world.

By way of an example, we will consider Flemming (2001b)'s analysis of stop voicing contrasts in English. Flemming (2001b) maintains that the voicing contrast in word-initial stops is realized primarily as an aspiration contrast, that is, underlyingly voiceless stops are realized as voiceless aspirated, and underlyingly voiced stops are commonly realized as voiceless unaspirated. Flemming (2001b) proposes the following analysis: a top-ranked constraint to maximize the number of VOT contrasts penalizes all solutions that neutralize the contrast. Surviving forms are analyzed with respect to two effort minimization constraints, one that penalizes initial voiced stops, and one that penalizes aspiration. A contrast between /t/ and /d/ in initial position would violate the former, higher-ranked constraint and therefore lose to the actually observed pattern, which violates only the constraint with the lowest ranking. This analysis, however, highlights several unsatisfactory aspects of the basic OT formalism: the ranking of the constraints appears arbitrary; Flemming (2001b) supplies no external evidence for the particular rank order of constraints, and if it were different, different patterns would be predicted. Moreover, the motivation of the constraints is based on intuition, rather than on actual measurements or model-based validation. The assumption that word initial stop voicing and aspiration require more effort than their absence may be plausible, but, again, no external evidence is presented to support these claims; (Flemming 2001b:47) acknowledges this with the tentative statement that "(a)spiration might be disfavored because of the effort involved, or because of the devoicing effect on a following vowel". Finally, "classical" OT with its strict ranking of discrete constraints is not suited to the analysis of continuous phonetic phenomena.

Some approaches in the OT tradition have combined phonological analyses with computational models of speech production, in order to allow for precise quantifications of effort and thus to obtain something better than purely speculative effort estimates. Kirchner (1998) offers an OT analysis of consonant lenition phenomena in the languages of the world that is based on H&H-compatible principles. Kirchner (1998) claims that a constraint termed $LAZY$ that minimizes effort, but is balanced against faithfulness and perceptual distinctiveness constraints, is responsible for lenition phenomena in speech. He develops a simple computational mass-spring model in order to obtain estimates of physical effort involved in different articulatory gestures. The mass-spring system represents an abstract articulator, which can be displaced in the direction of a boundary at an arbitrary location, representing the opposite vocal tract wall. Effort in this model is estimated as the force integral over a the time course of a "gesture", i.e., over the time interval during which the spring is displaced. The model utilizes an optimization algorithm in order to find the trajectory that minimizes effort for a gesture, given a user-specified spatial and temporal target.

As a concrete example, we may consider Kirchner (1998)'s account of spirantization, a putative phonological process by which stops are replaced by non-strident fricatives. Kirchner models stops by specifying a spatial target that actually lies beyond the vocal tract boundary, so that full closure is achieved, i.e., the model articulator achieves the displacement necessary to reach the vocal tract boundary, and is compressed against it. As for fricatives, Kirchner presupposes that the target lies shortly before the opposite vocal tract boundary, so that a narrow constriction is achieved. Moreover, he assumes that strident fricatives "require a relatively precise, sustained close constriction, in order to generate highly turbulent airflow" (Kirchner 1998:111), which is modeled by an antagonistic force active during the constriction interval of the gesture. Evaluation of the model yields the lowest energy cost for non-strident fricatives, whereas strident fricatives are even more "expensive" than stops. The non-strident version thus wins on the LAZY constraint in the OT evaluation and emerges as the overall winner. This, according to Kirchner (1998), accounts for the observation that unaffricated stops frequently lenite to non-strident but not to strident fricatives.

Kirchner (1998)'s modeling approach is a well-motivated attempt at quantifying articulatory effort, even though our above points about the somewhat arbitrary OT formalism apply to his work as well. An interesting feature of the mass-spring model is the utilization of an optimization algorithm for deriving the trajectory of the mass. The optimization problem is to find the energetically most efficient trajectory given a fixed spatial (and temporal) target, which is adjusted in the case of lenition. Thus, the model assumes that lenition implies the active readjusting of articulatory targets. This may come across as somewhat surprising; an arguably more intuitive view would be to interpret lenition as undershoot of an invariant underlying stop target. In modeling terms, this would require relaxation of the condition that the target *must* be reached. In order for this to be achieved, the model would presumably have to incorporate a (perceptual) cost on target undershoot directly in the optimization, rather than in a separate OT evaluation step. We will encounter an articulatory model conceived in exactly this way later in this chapter.

Boersma (1998) presents a somewhat similar computational OT framework based on efficiency-related assumptions. His basic tenet is that since language serves a communicative purpose, it is organized according to functional principles corresponding to the notions of minimization of effort and maximization of perceptual clarity. An interesting aspect of Boersma's work is that he aims at a more precise characterization of these principles. As for effort, he states that it "depends on at least six primitives: energy, the presence of articulatory gestures, synchronization of gestures, precision, systemic effort, and coordination" (Boersma 1998:149). Boersma introduces different constraints that implement these primitives, for example by penalizing articulatory displacement, speed,

duration, or the number of realized gestures. As for perception-related factors,Boersma posits minimization of perceptual confusion, minimization of categorization,[1] maximization of recognition (by using all available acoustic information on part of the listener) and maximization of information flow.

Boersma (1998) develops a fully-fledged articulatory synthesizer and a perception model based on psycho-acoustically motivated assumptions, which allow him to arrive at quantitative estimates of articulatory effort and perceptual clarity. Since the muscle forces in his articulatory model are conceptualized as mass-spring systems, articulatory effort is evaluated in a similar fashion as in Kirchner (1998)'s work: the effort estimate for a given model muscle is the force applied to it integrated over time, force being defined as the product of mass and acceleration. The effort computation also includes an estimate of the muscle's velocity, based on the reasoning that swifter movements are energetically more expensive. Moreover, Boersma (1998) assumes that holding a muscle in a non-resting position requires effort, which is incorporated by an "isometric contraction" constant. Boersma's approach also incorporates effort estimates that are not directly evaluated in model terms, such as the assumptions that gestural precision and coordination require effort. As for the perception model, Boersma (1998) recurs to findings on the psychoacoustics on speech perception to derive estimates of perceived spectrum, intensity and pitch – for example, his perception model non-linearly transforms the spectrum generated by the articulatory synthesizer into a "perceptual spectrum" hypothesized to reflect properties of the hearing system. These perceptual estimates are combined into a perceptual contrast metric that scores perceptual confusion probabilities based on perceptual difference limens established in psychoacoustic studies.

Boersma analyzes a multitude of empirical phenomena in his framework, which we cannot possibly discuss in their entirety. A single example, concerning the first formant (F1) value of the vowel /a/ in stressed and unstressed syllables, shall serve to illustrate the broad principle. Boersma (1998)'s perception model yields a confusion probability function, which increases with distance from the "prototypical" F1 value of a given vowel category. At the same time, the articulatory model yields an energy cost estimate that increases with jaw opening. Using a rough approximate transformation for translating jaw opening into F1 values allows Boersma (1998) to evaluate candidates with different F1 values. To this end, he splits the energy-minimization and the confusion-minimization constraints up into various sub-constraints penalizing increasing deviations from the acoustic target and increasing jaw opening, respectively. The sub-constraints

[1] (Boersma 1998:2) assumes that "in a world of large variations between and within speakers, the disambiguation of an utterance is facilitated by having large perceptual classes into which the acoustic input can be analyzed". Interestingly, this seems to be in direct conflict with Flemming (2001b)'s assumption that the number of perceptual contrasts is maximized.

are ranked in alternating fashion: [jaw opening > 4 cm] > [F1 ≤ 600 Hz] > [jaw opening > 3 cm] > [F1 ≤ 700 Hz] etc. Boersma (1998) assumes that in unstressed vowels, the constraint ranking is re-ordered such that the perceptual constraint is on the whole ranked relatively lower than in stressed vowels. This reproduces the frequent observation of vowel reduction in unstressed contexts in some languages, as the winning candidate would have a higher F1 with the stressed than with the unstressed constraint hierarchy.

In this particular example, the constraint ranking appears less arbitrary than in the Flemming (2001b) one, as the assumption of higher-ranked perceptual constraints in contexts associated with perceptual prominence is intuitively compelling. One interesting aspect of Boersma (1998)'s alternate constraint ranking, moreover, is that it turns the strict hierarchy into a more fine-grained scale. If this approach was taken further, the jaw displacement and the minimum F1 constraint could be turned into cost functions that could be evaluated in a fully continuous fashion, and Boersma (1998) also hints at this possibility. In what follows, we will devote our attention to studies that have followed this approach, replacing interaction of discrete constraints by continuous cost functions. In these approaches, constraint ranking is replaced by numerical weights that control the relative importance of the individual constraints. We will provide a fairly detailed discussion of these works, Flemming (1997) and Katz (2010), because their models focus on speech timing at the segmental level, and thus on a phenomenon that is very similar to what our model will be concerned with.

Flemming (1997) presents an approach of this kind which focuses on two phonetic phenomena, transitions of the second formant (F2) and duration ratios in consonant-vowel sequences. The crucial characteristic in both cases is contextual variation: vowel durations display a systematic (inverse) relationship with the duration of the following consonant. At the same time, consonant duration also co-varies inversely with vowel duration. A similar pattern is observed for formant transitions in CV sequences, which we already encountered in the discussion of Lindblom (1990)'s interpretation of Öhman (1966)'s data: F2 at the consonant is linearly related to the value of F2 at the steady state of the following vowel, a dependency that has been captured descriptively by so-called *locus equations*. At the same time, there is *target-locus proportionality*: F2 measured at the steady state of the following vowel also varies according to the F2 value at the consonant release. Consonant and vowel F2 thus stand in a relationship of mutual influence, which, according to (Flemming 1997:75) suggests the interpretation "that there are targets for the second formant of each consonant and vowel, [...] respectively, but these targets are systematically undershot with the actual F2 values being displaced towards each other".

Flemming (1997) proposes that this pattern is a consequence of the resolution of two constraints: one constraint, which can be related to the concept of faithfulness in OT, penalizes deviations from hypothetical targets. A second constraint penalizes fast movements, based on the assumption that formant trajectories can serve as a proxy for articulatory movements, and that fast movements require more effort than slow movements, all else being equal. This constraint thus incorporates the assumption that speakers strive to minimize effort, and could be interpreted as a markedness constraint, in that a production requiring more effort is considered more marked. Flemming (1997) implements these two constraints as mathematical distance functions. The faithfulness constraint comprises two terms, one that measures the distance between the actual F2 value at the release of the consonant, $F2_C$ and the hypothetical underlying production target, or *locus*, for this value, $F2_L$, and a second term that does the same for the F2 value at the steady state of the vowel, measuring the distance between the actual value $F2_V$ and the hypothetical target $F2_T$. The markedness constraint is simply the distance between the actual formant values $F2_C$ and $F2_V$: since Flemming (1997) makes the simplifying assumption that the duration of the transition is fixed, a greater distance between these values would require a faster movement. With this parametrization, the optimal resolution of the constraint system can be cast as a mathematical optimization problem: the actual formant values $F2_C$ and $F2_V$ have to be chosen so as to minimize the weighted sum c of the three (squared) terms:

$$c = w_c(F2_C - F2_L)^2 + w_v(F2_V - F2_T)^2 + w_e(F2_C - F2_V)^2 \tag{2.1}$$

It is obvious that, as long as $F2_L$ and $F2_T$ are not identical, the minimization of c will involve a trade-off: "the first two constraint terms will be minimized when $F2_C$ and $F2_V$ are equal to their target values, whereas the effort constraint is minimized when they are equal to each other" (Flemming 1997:76). The weighting factors w_c, w_v and w_e can be used to rank the constraints, analogously to the constraint hierarchy in OT. Crucially, however, this hierarchy is not strict: a candidate solution that incurs a somewhat larger violation of the top-ranked constraint than another solution can still be better if the other solution incurs a massively larger violation of the lower-ranked constraint. The optimization problem can be easily solved analytically for a given configuration of the weighting factors by taking the partial derivatives of c with respect to $F2_V$ and $F2_C$ and solving for the zeros in the resulting equations. Flemming (1997) shows how optimally solving equation 2.1 in this way reproduces the empirically observed trading relation between $F2_V$ and $F2_C$, with the relative weights determining how strongly both values diverge from their hypothetical targets.

The model obviously makes some strong simplifications – in addition to the assumption of fixed duration, Flemming (1997) notes that formant trajectories only provide a coarse

approximation of articulatory movements, and that supposedly different energy costs for different articulators are not directly taken into account. Moreover, it would of course be desirable for the model to account for the formant trajectory as a whole, not just for its static endpoints. In any case, abstraction is necessary in modeling, and Flemming (1997)'s technique offers a convincing and elegant account of the empirical data. Importantly, and in contrast to the purely descriptive locus equation accounts, Flemming's model takes an *explanatory* stance towards the phenomenon: the faithfulness and markedness constraints are directly related to securing information transmission (by faithful realization of acoustic targets) and minimizing effort, and thus constitute independent underlying principles in the sense of Lindblom and Engstrand (1989).

Flemming (1997) applies a similar optimization approach to a second acoustic-phonetic phenomenon, duration ratios in consonant-vowel sequences. in many languages, the duration of a vowel is found to vary inversely with the duration of the following consonant, and consonant duration simultaneously varies inversely with preceding vowel duration. Flemming assumes that this pattern is a consequence of constraint resolution between target durations of segments and larger constituents. He proposes the following model to capture this relationship:

$$c = w_{c1}(C1 - C1_T)^2 + w_v(V - V_T)^2 + w_{c2}(C2 - C2_T)^2 + w_\sigma((C1 + V + C2) - \sigma_T)^2 \quad (2.2)$$

Here, $C1, V, C2$ and $C1_T, V_T, C2_T$ are actual durations and hypothetical duration targets, respectively, for the onset consonant, the vowel, and the coda consonant in the target monosyllable, and σ_T is a hypothetical target duration of the syllable as a whole. Again, this quite clearly predicts a compensatory relationship: "E.g. if C2 is lengthened, V will shorten in compensation to prevent excessive violation of the syllable-level constraint, but the constraint on C2 duration will generally prevent total compensation" (Flemming 1997:89). Despite the identical form, equations 2.1 and 2.2 are conceptually different, however: the duration model 2.2 does not uphold the differentiation between faithfulness and markedness; rather, one would interpret it as a case of two conflicting faithfulness constraints at different linguistic levels. Consequently, there is also no consideration of effort proper in this model. We will provide a detailed critique of this approach below, after a discussion of the successive modeling work by Katz (2010).

Katz (2010) provides a detailed treatment of the intrasyllabic durational relationships already briefly addressed by Flemming. Katz (2010)'s experimental results inform a mathematical model that is based on the same principles, but appreciably more elaborate than Flemming's approach, and accounts very elegantly for the observed empirical patterns. Katz (2010) frames his account of intra-syllabic timing patterns in English in a broader theory of *compression effects*, which incorporates the basic principles of

Flemming's model 2.2: constraints on syllable and segment duration compete in speech production; hence vowel duration is shortened if consonants are added to a syllable. An important qualification raised by Katz (2010)'s results is that these effects vary with consonant manner. Katz reports a production study comparing vowel duration in syllables with zero, one or two consonants in either the onset or the coda, produced in uniform sentence contexts by six speakers of American English. An important feature of the analysis is the separate treatment of the duration of the steady states of vowels, and of the transitions between vowels and consonants. Results indicate that vowels are shorter in syllables with one than with zero consonants in either onset or coda, although effects sizes vary with consonant class. The situation is more complex for *incremental* shortening induced by adding a further consonant: additional shortening of a vowel if two consonants rather than one are included in either the onset or the coda is reliably observed if the consonant directly adjacent to the vowel is a liquid. With vowel-adjacent nasals, this effect is only observed if an additional consonant is added before an onset, but not after a coda nasal. With a vowel-adjacent obstruent, adding another consonant triggers no incremental shortening in either position.

Katz (2010) hypothesizes that this asymmetry is due to differences in *perceptual recoverability* of vowel information from consonants: liquids contain more cues to the identity of the adjacent vowel than obstruents and nasals. Therefore, speakers can "afford" to shorten vowels more strongly in the vicinity of liquids without jeopardizing perception, which is rewarded by a smaller violation of the hypothesized syllable duration constraint. This hypothesis is supported by a perception study: Katz (2010) reports an experiment on the identification of vowels from acoustically presented syllables as utilized in the production study, but with the vowel itself truncated or removed. Results indicate that subjects' performance in vowel identification co-varies with the type of consonant adjacent to the manipulated vowel, in a fashion that mirrors the durational patterns found in the production study: for example, truncated and removed vowels are identified better in syllables with vowel-adjacent liquids than in syllables with vowel-adjacent obstruents.

As a first pass, Katz (2010) proposes the following formal model to account for the timing relations within a simple consonant-vowel syllable:

$$c = w_1((d_x+d_t+d_y)-t_\sigma)^2+w_2((nd_y+md_t+ld_x)-t_x)^2+w_2((kd_x+jd_t+id_y)-t_y)^2 \quad (2.3)$$

This model is similar to Flemming (1997)'s equation 2.2, but appreciably more complex, as it incorporates the notion that segments are partially recoverable from acoustic information contained in adjacent segments. The first term penalizes the difference between the hypothetical target duration of the syllable, t_σ, and its realized duration, which is defined as the sum of the durations of the steady state of the vowel d_x, the transition d_t

and the consonant d_y. The two other terms penalize deviations of actual consonant and vowel duration from their respective targets t_x and t_y. As can be seen, realized duration terms for both segments (the inner brackets in the second and third quadratic term) are also functions of all three intrasyllabic constituent durations. Crucially, they are modified by the *recoverability coefficients* $i - n$, which implement the assumption that acoustic information from adjacent segments contributes to the recognition of a segment. Coefficients l and i, which represent the consonant- and vowel-internal acoustic cues, are set to 1. The other coefficients are set to values between 0 and 1, with the exact values reflecting the hypotheses about the amount of perceptual information the respective part of the syllable contributes to the perception of the vowel or the consonant, respectively. For example, k is set to a higher value for a liquid than for an obstruent, implementing the assumption that a liquid contains more acoustic information about an adjacent vowel than an obstruent. Katz (2010) shows that this model provides a convincing account of his experimental data for the case of *simplex* compression, i.e., the difference between zero and one onset or coda consonant.

Katz (2010) extends the model to cases where two consonants are added to a syllable onset or coda, in order to model incremental shortening effects. He reports that this model is not restrictive enough: it invariably predicts incremental vowel shortening, contrary to experimental results. For this reason, Katz introduces an additional assumption: he defines a minimum duration threshold for a given segment, by assigning arbitrarily high cost to segmental durations below a certain value. This idea is borrowed from the descriptive duration model by Klatt (1973), who hypothesized that segment durations in speech are characterized by *incompressibility*, i.e., a durational floor value beyond which they cannot be shortened. We will have to say quite a bit more about the concept of incompressibility in the course of this work. For the moment, suffice it to note that the introduction of this additional assumption allows Katz's model to reproduce the observed empirical patterns for incremental vowel shortening, with additional shortening if a second consonant is added to a vowel-liquid syllable, but not in the case of a vowel-obstruent syllable.

The models by Flemming and Katz offer intriguing accounts of the empirical phenomena under study. Katz's duration model in particular proves capable of predicting fine-grained temporal phenomena to a remarkable degree of detail, and his perception study of vowel recoverability from adjacent segments yields convincing independent motivation for his modeling assumptions. These studies thus provide powerful examples of how independently motivated assumptions can be utilized in rigorous mathematical models to derive empirically testable predictions. Nevertheless, some critical points may be raised. One underlying conceptual issue concerns the heavy reliance of this type of model on the notion of production targets that are apparently assumed to reside in

speakers' mental representations. While the assumption of targets that speakers attempt to realize is certainly reasonable for Flemming (1997)'s formant data, the idea that speakers represent *duration targets* for phonetic segments may be more debatable. The notion of duration targets for syllables in Katz (2010)'s model in particular appears to be somewhat awkward; Katz himself acknowledges that this constraint may be difficult to justify. As a possible motivation, he refers to a study by Quené and Port (2005), who reports temporal regularity of (stressed) syllable onsets to have a beneficial influence on spoken word recognition. Katz thus explicitly assumes that the syllable duration target constraint in his model is tantamount to the claim that syllable durations exhibit a tendency towards regularity. We think that this is not necessarily the case, however; granted that syllable duration targets of some sort exist, it may easily be conceivable that different syllables have different target durations. In any case, one may note that, whereas Flemming (1997)'s formant transition model exhibits a clear division of labor between economy- and clarity-related factors, this distinction is not so clear in the case of the duration models; as we said above, it is rather different types of faithfulness constraints that interact.

Katz's invoking of incompressibility in order to secure correct prediction of the incremental shortening results may come across as a somewhat ad-hoc device, but it is certainly a plausible and well-motivated one. He implements incompressibility as a hard floor, but also discusses a more elegant technique, whereby the faithfulness constraint is implemented as a hyperbolic function of deviation from the minimum duration. Katz does not implement this solution due to technical difficulties. In the course of this work, we will see how a somewhat similar modeling technique reproduces empirical findings on incompressibility, and that it can also be assigned independent motivation.

A final minor issue, which is also acknowledged by both authors, is that their models crucially rely on the quadratic form of equations 2.1 – 2.3 to derive the empirically observed patterns. The reason they give for this modeling decision is that the quadratic form ensures that costs grow quickly with constraint violations. While this is perfectly acceptable, an ideal model would include independent motivation not only of the general architecture, but also of the precise functions used to implement the individual constraints. For example, the cost functions that implement the faithfulness constraints could be designed so as to implement explicit assumptions about the perception of temporal or spectral properties of speech.

Despite these critical points, we would like to re-state our appreciation of the models by Flemming and Katz. Even if some of their assumptions may be better motivated than others, the success of the models in accounting for observed patterns is interesting in its own right, and definitely establishes their modeling approach as a promising research

direction. The model to be developed in this work is concerned with a similar phenomenon to the ones by Flemming and Katz, and it will be based on somewhat similar principles. We shall discuss commonalities and differences once we have introduced our own modeling approach.

2.2.3 Embodied Task Dynamics

We will now review the Embodied Task Dynamics (ETD) model of gestural sequencing in articulation (Šimko 2009). In our opinion, this model represents the best worked-out attempt at implementing optimality assumptions as an explanatory device for speech phenomena, and it will serve as the primary source of inspiration for our own model to be developed in the course of this work. In what follows, we shall provide a non-technical overview of the model architecture and discuss the results that ETD has achieved. For an in-depth discussion of the model architecture, see Šimko and Cummins (2010).

ETD is an extension of the task-dynamic implementation of Articulatory Phonology (Browman and Goldstein 1986, Saltzman and Munhall 1989), henceforth AP-TD. Articulatory Phonology assumes that articulatory gestures, as opposed to acoustically defined segments, are the primary phonological units in speech. AP-TD models these gestures as mass-spring systems, using the mathematical apparatus of linear second-order dynamics, based on the assumption that the behavior of the muscles acting upon the articulators is adequately described by the mathematics of dynamical systems. ETD introduces two crucial modifications to conventional AP-TD. The first is the embodiment: in AP-TD, the mass of the articulators is disregarded and set to unity for all articulators. By contrast, ETD crucially relies on the assumption that different articulators have different masses. ETD also takes into account collisions between articulators and the boundaries of the vocal tract, and, moreover, explicitly assumes that all articulators are anatomically linked and thus mutually influence each other, whereas AP-TD abstracts away from this property of speech and models individual gestures as context-independent. The second modification concerns the specification of the temporal sequencing of individual gestures. This is done either manually or by using techniques such as neural networks in AP-TD. In ETD, this is where optimization comes in: the temporal sequencing of gestures is specified using an optimization algorithm that, for a given utterance, determines the optimal sequencing and realization of gestures with regard to a parametric cost function, encompassing component costs related to measures of articulatory effort, perceptual clarity and overall time, as explained below.

At the core of ETD is a highly simplified model of the human vocal tract, as shown in the left panel of Figure 2.2: masses representing the jaw, tongue body, tongue tip, upper

and lower lip are attached to hard boundaries by springs, as shown in the Figure. The model thus ignores most of the details of human vocal tract anatomy, but it captures the basic anatomical facts of the modeled articulators: the tongue body is attached to the jaw, and the tongue tip, in turn, to the tongue body; the lower lip is also attached to the jaw, but independently of the tongue, whereas the upper lip is attached to the maxilla. The different sizes of the black circles in the Figure represent the different masses of the articulators – for example, the jaw has a greater mass than the tongue tip. This model is capable of representing differences in the vertical position of the tongue body – i.e., differences in vowel height – as well as alveolar and bilabial closures.

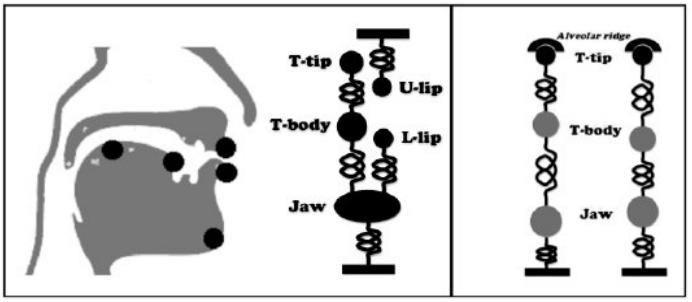


FIGURE 2.2: Left panel: schematic overview of the ETD model (reproduced from Šimko and Cummins 2011). Right panel: illustration of different model configurations that lead to the same constriction.

In ETD as well as in AP-TD, any articulatory movement is instantiated by an underlying *task* specifying the spatial goal of the movement. These tasks are encoded in terms of the values of *tract variables*, parameters describing the location and degree of relevant vocal tract constrictions. The movements necessary to reach the articulatory targets are described by linear second-order dynamical systems equations, as shown in equation 2.4

$$M\ddot{z} = -K(z - z_0) - B\dot{z} \tag{2.4}$$

Here, M refers to the mass, K to the *stiffness* and B to the *damping* coefficients, z_0 to the target positions of the tract variables, and z, $\dot{z}$ and $\ddot{z}$ refer to position, velocity and acceleration, respectively, of the articulators involved. The stiffness parameter is related to the resistance of the spring to external forces acting upon it, with higher stiffness values leading to a more resistant spring and, hence, swifter movements. The damping parameter is related to the nature of the oscillatory behavior of the spring. In articulatory modeling, the value of this parameter is usually set to a certain value analytically related to mass and stiffness that is referred to as *critical damping*, which ensures that the movement smoothly approximates its target and does not continue to oscillate after the target has been reached (Šimko 2009).

The specification of tasks in terms of constrictions (rather than absolute positions of all articulators involved) has an important consequence: there are many different vocal tract configurations that will result in the same constriction, as shown for the example of an alveolar closure in the right panel of Figure 2.2. The problem that the model faces, then, is to select among these possible alternatives. In technical terms, the model, for a given gesture, has to select an *activation interval*, i.e., the temporal interval during which the dynamic that instantiates the gesture is active, and a stiffness value for the gesture. This *degrees-of-freedom problem* of how to choose from a potentially vast space of alternative possibilities to achieve a certain behavioral goal is thought to be a pervasive characteristic of speech production, as well as other types of coordinated movement (Lindblom 1999).

The basic hypothesis of ETD is that this problem is solved by optimization: articulatory gestures are shaped and coordinated so as to optimally satisfy the requirements of minimizing articulatory effort and maximizing perceptual clarity. This is implemented using a parametric cost function that quantifies the degree to which a given rendition of an utterance satisfies these two requirements (as well as a third one, as introduced below), and an optimization algorithm that finds this optimal gestural score. Weighting factors allow for specifying the relative impact of the individual components of the cost function. The parametric cost function is defined as in Equation 2.5:

$$C = \alpha_E E + \alpha_P P + \alpha_D D \tag{2.5}$$

Component E is related to articulatory effort. The framework of mass-spring equations allows for a straightforward parametrization of effort in ETD: it is computed by summing all forces that act upon the model's articulators during the realization of a simulated utterance, the force for an individual articulator being defined as the product of its mass and its acceleration. Articulatory effort is thus causally related to the stiffness of the movement, as higher stiffness will result in swifter movements. There may be alternative definitions of effort in speech, but this metric makes intuitively plausible predictions: all else being equal, moving a heavier articulator will be more costly than moving a lighter one, and stiffer movements will also be more costly than less stiff ones.

The *parsing cost* P in the model represents the hypothesized impetus towards maximizing perceptual clarity. It comprises two sub-costs: first, Šimko (2009) assumes that failure to fully reach articulatory targets will result in acoustically degraded output and impede perception. This is captured by a *gestural precision estimate*, which for a given tract variable is inversely proportional to the difference between its target and its actual value. The gestural precision estimate for a given articulator is normalized to the difference between the target and the resting position of that articulator, so that inherently smaller movements are not evaluated as inherently more precise, and may thus take on

values between 0 and 1. Second, (Šimko 2009) assumes that longer realizations of gestures will facilitate perception. This is implemented by a *temporal realization estimate*, which monotonically increases with the duration of a gesture's realization in the interval $[0, 1]$. Gestural precision estimate and temporal realization estimate are combined in a single parameter, *realization degree* r_g, by multiplying their respective maxima for a given gesture. The parsing cost P_i for a given gesture i, then, is defined as $P_i = 1 - r_{gi}$ (recall that r_g has an upper bound of 1), and the parsing cost for an utterance is the sum over the parsing cost terms for all the gestures in the utterance.

Before going further, we shall provide some more details on the temporal realization estimate, $d_g(t)$. Šimko assumes that this estimate is not a linear function of the duration of a gesture, but: "that it increases dramatically within a few first tens of milliseconds after the onset of gestures [sic] prominence interval, and then remains virtually unaffected." (Šimko 2009:137). The particular function used given in Equation 2.6:

$$d_g(t) = \frac{2}{\pi} arctan(c(t - t1_1)) \tag{2.6}$$

The t and t_1 parameters in Equation 2.6 refer to the start and end point of a gesture, respectively, and c is a constant that can be used to adjust the slope of the function. Resulting trajectories for some values of c are shown in Figure 2.3.

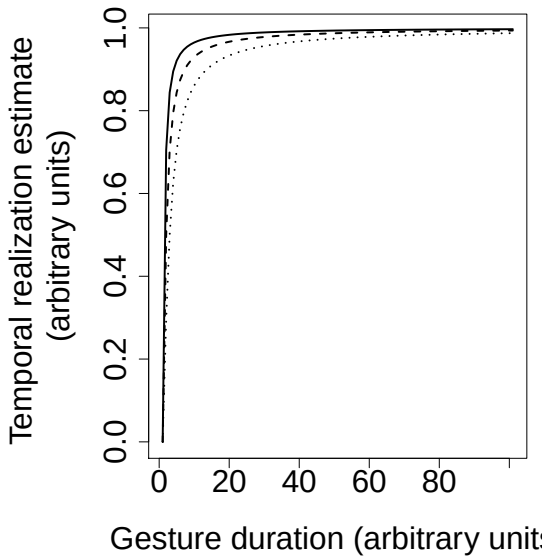


FIGURE 2.3: Plot of the temporal realization estimate in ETD for $c = 2$ (solid), $c = 1$ (dashed) and $c = 0.5$ (dotted).

Is $d_g(t)$ a realistic estimate of temporal realization of a gesture? Šimko and Cummins (2011) cite Gray (1942) in support of this modeling decision. Gray (1942) presented subjects with isolated vowels of varying duration and recorded the number of correct identifications. Inspection of his numerical results does indeed tentatively suggest a

function similar to $d_g(t)$. In this thesis we will discuss some more recent results from speech perception research that support Šimko's modeling of the temporal realization estimate for larger prosodic units, and an integral part of our own model to be developed in the course of this work will rest on this argument.

The model components E and P implement the hypothesized trade-off between tendencies towards hypo- and hyperarticulation. In addition, Šimko (2009) introduces a temporal dimension to the model by invoking the third component, D. He motivates this by the observation that the H&H continuum is to some extent independent of changes in speaking rate: speakers can, within certain limits, speak quickly and still articulate carefully, or, conversely, articulate both slowly and imprecisely. Component D, which is simply a linear function of the time interval between the onset of the first and the offset of the last gesture of a simulated utterance, implements this temporal dimension, providing an independent control mechanism for overall speaking rate. The parameters α_E, α_P and α_D in Equation 2.5 are scalar weighting factors that can be used to manipulate the relative importance of the three components. For example, a relatively high value of α_P would simulate conditions that favor hyperarticulation, such as speaking in a noisy environment, or to a non-proficient listener. Given a specification of a sequence of gestures and a set of values of the weighting factors, the problem the model has to solve is to find activation intervals and stiffness coefficients for the individual gestures such that cost function C is minimized. This is implemented in ETD using simulated annealing, a standard method for solving nonlinear optimization problems, which is roughly similar to the algorithm utilized by Liljencrants and Lindblom (1972) as described above.

Figure 2.4 shows results of an ETD simulation of two utterances, /abi/ (left plot) and /iba/, i.e., sequences of two vowels of different height flanking a bilabial stop (arbitrarily defined as voiced). The upper panels show the optimal *gestural scores*, i.e., the temporal sequences of activation intervals for the three gestures in both utterances. The lower panels plot the trajectories of the jaw (thick solid line), tongue body (thin solid line) and lips (dashed lines) over time. Two observations can be made on this plot. First, it shows that the model reproduces a fundamental property of speech at the articulatory level, namely the fact that gestures are co-produced, i.e., that there is temporal overlap of consonantal and vocalic gestures. Šimko (2009) argues that this result is non-trivial, given that both types of gestures exert simultaneous and sometimes contradictory pressures on articulators. Moreover, optimization in the model is initialized with gestural scores with no overlap. Co-production, according to Šimko (2009) thus emerges as the optimal way of coordinating consonantal and vocalic gestures given the articulatory, perceptual and temporal constraints acting upon the system.

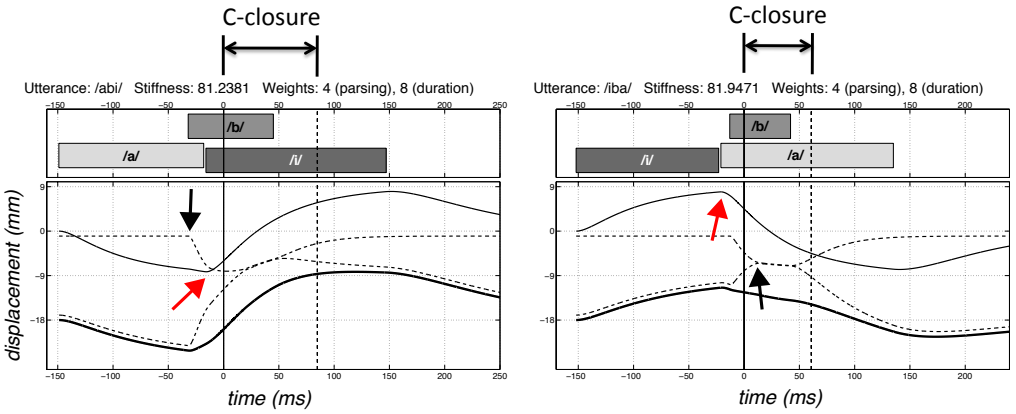


FIGURE 2.4: Plot of /abi/ (left) and /iba/ simulations in ETD (reproduced with modifications from Šimko and Cummins 2011). See text for details.

Closer inspection of Figure 2.4 reveals a more interesting result: one may note that for the /abi/ sequence, the onset of the closing movement of the lips (red arrow) precedes the onset of the tongue body movement towards the /i/ (black arrow), whereas in the /iba/ sequence, it is the tongue body movement that starts first, before the lip movement. This asymmetry is precisely what Löfqvist and Gracco (1999) have found in an articulatory study of VCV sequences in Swedish. Šimko and Cummins (2011) report that this pattern is reproduced across a range of parameter settings. They argue that their embodied optimization account offers a straightforward interpretation of the effect: "Given that the lips are closer together during production of an /i/ vowel than an /a/ vowel, this makes sense. Movement towards closure of the jaw, and hence also the tongue body, can start later for a medial bilabial consonant uttered after /i/ than after /a/, as the distance to be traversed to the point of consonantal closure is smaller" (Šimko and Cummins 2011:552). Another interesting finding is that the *phasing*, i.e., the relative timing of the onset of the second vocalic gesture with respect to the consonantal gesture is remarkably stable over a range of parameter settings, whereas the relative timing between the onset of the first vowel and that of the consonant varies widely. Šimko and Cummins (2011) interpret this outcome with regard to the cross-linguistic and developmental preference for CV over VC structures, arguing that it points towards the emergence of syllable structure from the model's optimization principles.

Subsequent studies using ETD have provided further interesting results. Šimko et al. (2014b) investigate the contrast between singleton and geminate consonants in Finnish, modeling gemination by *locally* increasing the parsing cost weight α_P for the medial bilabial consonant gesture in a VCV sequence. This leads to a longer closure duration, which is of course not surprising. However, the authors report an interesting observation on the relative phasing between the consonantal and the following vocalic gesture in the sequence: the duration of the inter-onset-interval between both gestures varies with

linearly increasing α_P in a highly nonlinear fashion, exhibiting a sudden quasi-discrete jump between two relatively stable plateau regions, in which relative phasing is not strongly affected by α_P variation. The authors interpret this result as showing that discrete phonological contrast, in this case between singleton and geminate consonants, may emerge from optimization over continuous dynamical parameters.

In a further study, Beňuš and Šimko (2014) investigate the emergence of prosodic boundaries in speech under continuous variation in speaking rate and articulatory precision. They analyze VCV sequences occurring after a syntactic clause boundary from two sentences produced by four speakers of Slovak, under the instruction to produce these sentences with continuously varying speaking rate and articulatory precision. Beňuš and Šimko (2014) observe that, as speaking rate is reduced, speakers introduce prosodic boundaries before the test sequence. The strength of these boundaries also correlates with rate; weaker boundaries marked by glottalization are introduced at medium rates, while even stronger slowing also triggers the occurrence of silent intervals. A particularly interesting result is observed on the relative phasing between the consonantal release before the boundary and the onset of the medial consonantal gesture in the VCV sequence: this relative phasing exhibits a quadratic relationship with speaking rate; the medial (bilabial) consonant in the VCV sequence is more in-phase at medium than at fast or slow rates. Beňuš and Šimko (2014) report simulations using a modified version of ETD that zooms in on modeling lip aperture and ignoring the rest of the vocal tract architecture, modeling the prosodic boundary by locally relaxing the duration cost weight α_D at the boundary location. They show that inducing continuous rate variation in the model simulations (by means of *globally* varying α_D) reproduces the qualitative pattern of results observed in the real data: the relative phasing of the medial consonantal gesture varies in a non-linear fashion as a function of overall speaking rate.

These results are quite interesting, because they highlight the potential of local parameter variation in the model. In these cases, some questions may be warranted about the motivation of the particular modifications – how does "a local increase in perceptual clarity demands" represent gemination, and likewise, what is the independent motivation for modeling final lengthening by locally adjusting the speaking rate parameter – but the results are interesting enough by themselves. Local variations in the weighting parameters that govern H&H scale variation will feature importantly in our own model to be presented in this work.

An important principle emphasized in Šimko (2009) and subsequent work is that ETD with its computationally costly optimization of dynamical parameters is not to be viewed as a real-time production model of speech. Rather, (Šimko 2009:149) hypothesizes that

"the trade-offs captured by our cost function definition play their role during the development of speech as a skilled human activity, and are thus reflected in the phonological laws underlying speech production". Whatever the interpretation, modeling results suggest that the trade-off between minimizing effort and maximizing perceptual clarity as implemented in ETD offers an intriguing account of various gestural coordination phenomena in speech. In particular, the embodied nature of the model provides physiologically plausible explanations for the modeled phenomena. The underlying assumption that the degrees-of-freedom problem in the coordination of speech gestures also provides a unifying link with other successful models from non-speech domains. Of course, ETD is highly simplified in many respects, but the fact that it still makes realistic predictions is all the more encouraging. For these reasons, the ETD model will serve as the primary source of inspiration for the design of our own model to be developed in this work.

2.2.4 Other Approaches

One line of research that is closely related to the approaches discussed so far investigates the relationship between measures of predictability or information density in language and acoustic-phonetic parameters of the speech signal. These studies generally support the conclusion that speakers strive to transmit information in an efficient manner: in regions of the speech signal that convey important information, great care must be taken to secure successful transmission of these regions; conversely, in parts of the speech signal that are informationally redundant or predictable from the context, speakers can afford to be less concerned about the demands of the listener. These opposite poles are thus readily interpretable in terms of hyper- and hypoarticulation. In a classical study of this kind, Lieberman (1963) had three speakers of English read sentences with target words varying in predictability from their contexts, such as *nine* in *A stitch in time saves nine* (high predictability) versus *The word that you will hear is nine* (low predictability). He reports higher durations and amplitudes for target words in the low-predictability condition and concludes that "in connected fluent speech the speaker calls attention to the words that he thinks are non-redundant" (Lieberman 1963:181).

The study by Aylett and Turk (2004) provides a more recent example of this way of reasoning. The authors frame their investigation of the relationship between speech timing and language redundancy in an information-theoretic account of speech production and processing. They argue that "the drive for speakers to achieve robust information transfer in a potentially noisy environment while conserving effort" (2004:32) predicts *smooth signal redundancy*, i.e., a negative correlation between syllable duration and the predictability of syllables, so as to ensure that syllables with a low predictability – and, thus, a high information load – are successfully transmitted. The idea that effort is

conserved plays a crucial role in this argument: "another way of ensuring that elements with low levels of language redundancy are recognized correctly would be for speakers to produce all elements in an utterance with maximal duration and clarity. However producing speech in this way would be highly inefficient compared to producing more predictable elements with reduced duration and effort" (2004:33).

Aylett and Turk (2004) investigate their hypothesis in a corpus of task-oriented dialogues in Glasgow English, by regressing syllable durations on three measures of language redundancy, word frequency, syllable trigram probability and givenness, the latter operationalized as the number of times the referent of the word that a syllable is part of has been mentioned before. Controlling for prosodic prominence and boundaries, they find that, as predicted, all three measures exhibit significant negative correlations with syllable duration. Investigating the relationship between redundancy, duration and prosodic prominence further, the authors find that the durational variation explained by language redundancy and prosodic structure overlap to a large extent, i.e., prosodically prominent syllables also tend to be non-redundant. They conclude that prosodic structure in English is in fact primarily used to modulate linguistic information flow. Prominence marking is thus interpreted as a means to secure the transmission of low-redundancy items, by making them acoustically more salient.[2] Results from further studies (e.g. Aylett and Turk 2006, Baker and Bradlow 2009, Jurafsky et al. 2001, Pluymaekers et al. 2005, Samlowski et al. 2013, Van Son and Van Santen 2005) support the conclusion that efficient information transmission plays an important role in speech production, by showing that measures of language redundancy correlate with temporal, and also with spectral reduction in speech. If it is taken for granted that these reduction processes reflect speaker strategies for saving effort, these findings provide a strong argument for the plausibility of the assumption that trading off between minimizing effort and maximizing communicative success play a decisive role in shaping speech patterns.

Howard and Messum (2011) present Elija, a computational model of speech acquisition in infants that makes use of optimization principles. It is implemented as a production-perception loop, featuring a "vocal tract" based on an articulatory synthesizer and a simple speech recognition algorithm that allows the model to perceive speech. Elija starts speech acquisition by "babbling", that is, initializing the articulatory synthesizer with control parameters determined by optimization, as explained below. During this process, Elija forms categories, by clustering produced sounds in the acoustic and articulatory space. In the following learning phase, the caregiver – a human experimenter – trains the model by repeating those sounds that resemble actual human speech sounds, which

[2]This view is somewhat relativized by Turk (2010), who states that facts such as the existence of language-specific stress assignment rules argue against viewing prosodic prominence structure as being entirely explained by information density modulation.

the model then utilizes to tune its parameters so that its productions gradually become more adult-like. Howard and Messum (2011) report that they were able to teach the model to produce a variety of English-like sounds and CV sequences, and even a number of simple English words.

The optimization-based babbling routine in Elija works as follows: the model selects those vocal tract actions that minimize a weighted sum of measures of production effort and acoustic salience. Effort is estimated based on cost metrics that score the actions of the articulatory synthesizer and salience is evaluated based on analysis of the acoustic output as "perceived" through Elija's own microphone, based on features such as acoustic power and high to low frequency ratio. The optimization also comprises a further component, "diversity", which penalizes similarity of a vocal tract action to previously tried vocal tract action, providing an impetus for the model to try out a variety of vocal tract actions. Interestingly, Howard and Messum (2011) report that many of the sounds Elija discovers in the (unsupervised) babbling phase do already resemble actual English vowels, which lends plausibility to the employed optimization assumptions.

We will discuss the effort estimate in some more detail here. It comprises two components, articulatory and voicing effort. The former is proportional to the velocities of all articulators involved in a vocal tract action, based on the reasoning that fast movements are energetically expensive. The second term is designed to serve as an estimate of the effort necessary to sustain phonation over time. This is suggested by a plot of the two effort components for a short utterance as produced by Elija, as shown in Figure 2.5:

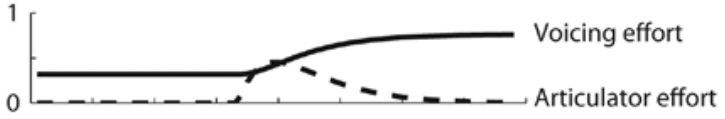


FIGURE 2.5: Articulatory (dashed line) and voicing effort (solid line) over time as computed by the Elija model for the utterance /tiː/ (reproduced from Howard and Messum 2011).

The "voicing effort" component is a particularly interesting feature of Elija. Discussions of effort in speech are often narrowly focused on energy costs of articulatory movements. However, we think that once one accepts the conclusion that minimization of effort is a relevant parameter in speech production, the metabolic energy needed to sustain phonation over time is in fact a quite natural thing to consider. This is especially true for a model that looks at speech in terms of larger time scales, such as the one we are going to introduce in this thesis. Consequently, Howard and Messum (2011)'s technique of splitting effort into an articulatory and a phonatory component will be taken into consideration in the design of our optimization-based model of speech timing.

Kochanski and Shih (2003) describe Stem-ML, a quantitative intonation model that incorporates assumptions compatible with those of H&H theory. Stem-ML models intonation contours using an overall phrase curve, onto which local accents are superimposed. These accents come with *tags*, containing specifications of the tonal target and shape of the accent, and also with specific strength values, which are related to their semantic or pragmatic importance in their respective utterance context. The actual F0 trajectory that is predicted is a product of optimization: Kochanski and Shih (2003) assume that speakers trade off maximization of communicative success and minimization of effort in the realization of intonation contours. As in the models already described, this is implemented using and optimization algorithm that minimizes the weighted sum of an effort-related and a perception-related term. Effort in Stem-ML is primarily related to the smoothness of the intonation curve, based on the assumption that a non-smooth F0 trajectory will put high strain on the muscles controlling the vocal folds, as they have to perform fast and strongly accelerated movements in such circumstances. Perceptual clarity is conceptualized as an error term that penalizes deviations between the actual F0 contour and the target heights and shapes specified by the accent tags. The strengths of the accent tags function as optimization weights: high strength values assign high costs to deviations from target and shape specifications, whereas for accents with low strength, it is more acceptable to undershoot F0 targets, so that effort can be saved.

While Stem-ML incorporates physiologically and cognitively plausible assumptions, its main purpose is to provide quantitative descriptions of intonation contours, and not so much to derive principled explanations for qualitative patterns. Kochanski and Shih (2003) show that their model delivers close approximations of tonal contours in Mandarin Chinese, but this is achieved by explicitly fitting the model parameters to the data. Yet, Stem-ML does seem to capture some general qualitative patterns of tonal contours. Kochanski and Shih (2003) discuss two such effects, tonal coarticulation in terms of the height and the shape of tones. In connected speech, both parameters are typically affected be the environment of a tone, manifested in effects such as the lowering of a high tone in a low neighborhood, or pitch range compression, compared to tones produced in isolation. Kochanski and Shih (2000) state that such effects are borne out by the optimization assumptions implemented in their model: smoothness constraints and perceptual constraints interact and give rise to trading relations between neighboring tones. Moreover, while empirical tests of Stem-ML seem to have focused on Mandarin Chinese, a tone language, Kochanski and Shih (2003) argue that the supposed universality of optimization assumptions should earn the model cross-linguistic applicability.

2.2.5 Criticisms of Efficiency-Based Explanations in Speech Science

In the Introduction to this work, we already briefly touched upon the question whether the assumptions of H&H theory and related approaches are realistic. In what follows, we shall address commonly phrased criticisms of efficiency-based explanations of speech patterns. One such objection, portrayed by Pouplier (2012) (without adopting this position), is based on the question why not all languages are the same if they are the product of efficiency constraints, which are supposedly universal. This argument may initially seem plausible, particularly with a view to the optimization models of vowel systems discussed earlier in this chapter: why are there some languages with apparently non-optimal vowel systems? As a basic response, however, one may note that the argument seems to assume that the current state of the languages of the world represents the end point of language evolution. In reality, language evolution is an ongoing process, and the current situation cannot be evaluated as if it marked the end point of the phylogenetic development of the world's languages.

More importantly, if language evolution is envisioned in terms of optimization, the real optimization problem posed by the evolution of a language over time is by orders of magnitude more complex than simulation approaches centered on isolated subsystems, such as vowel inventories or gestural coordination, are able to convey. Complete representation of the hypothetical optimization problem implied by language evolution would require considering an enormous variety of language-internal and external factors. Our discussion so far has revealed that even much simpler optimization problems may entail the possibility of converging onto solutions that are only locally optimal, and one may assume that this is even more likely for an optimization problem as complex and multidimensional as language evolution. On this view, different languages could be conceived as corresponding to local minima of the cost landscape, onto which linguistic communities have converged through repeated interaction. The most important point, finally, is that the putative optimization problem that is language evolution incorporates social and cultural factors, which include those that directly counteract convergence between different linguistic communities, such as the desire to express group identity and distinction from other social groups through language use. The language variation as a function of social class in English presents a prime example. The argument that economy principles predict sameness of all languages is simply naïve and simplistic.

A more serious criticism of efficiency-based approaches in speech research concerns their involvement of effort as an explanatory device. Physical effort in speech is difficult to quantify, and no single agreed-upon measure exists. Pouplier (2012) offers a nuanced critique of the concept, arguing that commonly employed metrics, such as the number of gestures or the distance traveled by an articulator, are overly simplistic. Pouplier points

out that invoking effort as an explanatory device for phonological processes or language change in particular entails the danger of circularity if the mere observation of a process is taken as an indication of minimization of effort without supplying external evidence. She argues that lenition phenomena in particular reflect skill, rather than "laziness" in speakers and may be deliberately used for information structure purposes. Pouplier argues that different speaking styles are not more or less optimal, but equally efficient in the contexts in which they are used. Moreover, various researchers have questioned the proposition that conservation of effort plays any role in shaping speech, arguing "that energetic considerations are at best marginal, as the relatively slight masses are acted upon by disproportionately powerful muscles". (Šimko and Cummins 2011:531, referring to an argument in Keller 1987).

These objections need to be taken seriously. We definitely agree with Pouplier (2012) that realistically estimating effort in speech is complex and multidimensional. Pouplier (2012) underscores her point with evidence from tongue body movements in intervocalic velar stops, which are characterized by trajectories that would be classified as non-optimal if the distance traveled by the articulator were to be used as the relevant effort metric. Pouplier states that the movement trajectory becomes explicable once biomechanical properties of the tongue are taken into account. Following Šimko et al. (2014b), however, we would argue that this does not constitute evidence against effort-based explanations per se, but rather for using the *right* effort metrics. Based on results from Perrier et al. (2003), Pouplier concludes that the observed trajectories result from "the tongue muscle orientation and activation patterns and the interaction of the tongue with the hard palate" (2012:154). We think it quite reasonable to assume that tongue movements directed against tongue muscle orientation are energetically suboptimal on measures other than the distance traveled by the the tongue body. This is, in fact, quite compatible with the notion of articulatory synergies, which is often invoked in efficiency-based accounts of speech patterns (Lindblom 1983, Šimko 2009). Pouplier (2012) also contends that different speaking styles are equally efficient in the contexts in which they are used. We fully agree with this statement, and find it rather surprising that Pouplier seems to view it as an argument against economy-based explanations in speech science: the notion that a speaking style that requires more effort than another may still be more efficient in certain situations, for example if greater clarity is required, in fact constitutes the very essence of such explanations; what they claim is precisely that speech production is *efficient*, not that it is *effortless* in absolute terms.

In the absence of external evidence, we cannot directly refute the claim that energetic considerations play no role in speaking due to the disproportionate relationship between muscle forces and articulator masses. The same goes for Pouplier (2012)'s hypothesis

that reduction phenomena in speech are not a consequence of "laziness", but are actually actively employed by speakers for structuring information. Yet, the existence of implemented models that successfully account for speech phenomena based on efficiency principles argues for the relevance of these principles, even though models can of course deliver only sufficient (and not necessary) proof. It is, in fact, a primary function of computational modeling to assess the influence of factors that are not directly observable, such as effort in speech. The model we are going to introduce in this work will contribute further evidence for the plausibility of efficiency-related explanations in speech science.

2.3 Discussion

We have seen that efficiency-related approaches provide possible explanations for empirical findings from a variety of sub-disciplines of speech research. This, in our opinion, establishes the assumption that speech patterns are shaped by trade-offs between demands to minimize effort and maximize perceptual clarity as a promising explanatory device. In particular, the success of existing implementations of optimization-based models encourages us to apply this methodology to the domain of suprasegmental speech timing, for which currently no such implementation exists.

Importantly, H&H theory and related approaches possess the potential to provide us with genuine *explanations* for phenomena to be modeled, as these approaches are based on independent principles that operate at a more basic level than the linguistic categories we will be concerned with. This is most apparent from the treatment that phonological structure receives in H&H theory: it is thought to *emerge* from constraints on speech production and perception, rather than representing arbitrary structure that hedges essentially random phonetic variation, as formalist approaches suggest. The appeal of this functionalist stance towards speech, ultimately, comes from its compatibility with more general theories of biological evolution. On this view, speech production represents an adaptation to environmental conditions, and the speech forms that are phonologized by linguistic communities are conceivable as the "survivors", which are best adapted to the environment they are used in.

One core insight from our discussion, finally, is the imperative that theoretical proposals be implemented, so that they can be rigorously tested. The critical treatment by Pouplier (2012) in particular underscores this point, highlighting the fact that theoretical reasoning alone is not sufficient to determine the suitability of a proposed mechanism and may result in inadequate explanations and circular reasoning. As stated above, we have identified the approach utilized in Šimko (2009) and related work as a particularly apt methodology for implementing efficiency principles, and we will consequently adapt

it to the modeling of suprasegmental speech timing phenomena. A concise review of such phenomena is provided in the subsequent chapter of this work.

Chapter 3

Suprasegmental Speech Timing

3.1 Introduction

In this chapter, we shall present a concise review of results from research on suprasegmental speech timing, in order to obtain an overview of the phenomena that need to be accounted for by our model. As for the definition of the term "suprasegmental", we adopt a formulation from (White 2002:6): suprasegmental timing effects are "those that arise from the linguistic structure of a syllabified string". White actually uses this definition to denote *suprasyllabic* timing effects, while he refers to *suprasegmental* timing effects as "those which result from the organisation of segments into a string of syllables" (2002:ibid.), including primarily lexical stress. We will conveniently subsume both classes of effects under the cover term "suprasegmental", as lexical stress is commonly thought to apply to syllables, and we presume that it is perceptually established by comparing the prominence levels of syllables in a sequence, alluding to Lehiste (1970)'s classical definition of prosody. Quoting Klatt (1976), (White 2002:ibid.) states that suprasyllabic (suprasegmental, in our terminology) timing effects fall into three categories: "boundary-related lengthening, lengthening due to prominence, and shortening due to the phonological size of the constituent". In this chapter, we will review cross-linguistic evidence for these three classes of effects. Separate attention will, moreover, be devoted to interactions of these effects with changes in overall speaking rate due to different external conditions.

Our usage of the term "suprasegmental" may roughly be glossed as describing those timing effects that apply at the syllabic level, and the model we are going to introduce later will represent syllable durations. Many of the results we review in this section are actually stated in terms of vowel rather than syllable duration. Yet, their common characteristic, following from our above definition, is that they are hypothesized to

be consequences of the suprasegmental organization of speech. We assume that by and large, suprasegmental timing processes should have comparable effects on the net duration of a syllable than on the duration of its nucleus, even if they may affect other parts of the syllable to different degrees, or not at all.

Finally, it is important to note that the above categories of suprasegmental effects do not apply to the description of all languages of the world. They are, in particular, tailored to the description of *stress-accent languages*, which are defined precisely by their usage of linguistically conventionalized prominence contrasts for expressing syntagmatic relations between linguistic units (Hyman 2006). These categories do, for the most part, not apply to *tone languages*, which utilize acoustically salient dimensions (in particular F0) that would be employed for syntagmatic purposes in stress-accent languages mainly for signaling paradigmatic contrasts between lexemes (Beckman 1986, Hyman 2006, Jun 2005).[1] This review will consequently exclude purely tonal languages. Apart from this distinction, we do not attempt at strict separation, however; for example, we will consider evidence from French and Korean, which supposedly have neither stress nor tone (Hyman 2006). Moreover, stress accent languages may or may not have lexical pitch accents in addition to syntagmatic prominence relations (for example, Swedish versus English, cf. Hyman 2006), but this will not concern us here. In any case, our review cannot possibly be exhaustive cross-linguistically, and it will necessarily be biased towards the well-researched languages, in particular English. We thus cannot claim that the results we present are universals, but it should be possible to point out cross-linguistic tendencies, and we will consider evidence from various languages wherever available.

3.2 Prominence Effects

3.2.1 Introduction

When we listen to a speech utterance, some of its parts seem to us to "stand out" more than others. This phenomenon is commonly referred to as *perceptual prominence*. Following Wagner (2002), we define it as the gradually perceived strength of a syllable or a larger prosodic unit relative to its environment. Although this definition makes prominence an inherently perceptual phenomenon, a large body of research has shown that it correlates robustly with acoustic parameters of the speech signal: prominent units in speech tend to be marked by enhancement of various intonation-related parameters,

[1] This explanation simplifies things somewhat. For example, prominence relations between words may be expressed by prosodic means even in tone languages (Xu 1999). In any case, we want to exclude tone languages from consideration because we believe that any satisfactory model of these languages must take tonal (i.e., F0) properties into account, whereas our model will exclusively focus on timing.

increased vowel space, intensity, and, crucially for the present study, duration (e.g. Fant and Kruckenberg 1989, Fry 1958, Heuft et al. 2000, Streefkerk 2002). These parameters may be employed for signaling prominence to different extents in different languages and contexts. For the present discussion, it will be sufficient to note that increased duration has been established as a reliable correlate of prominence in various languages (cf. references in Wagner 2002).

In stress-accent languages, some prominence distinctions have become conventionalized in the linguistic system. One such distinction is *lexical stress*: in these languages, every content word typically contains one dominantly prominent syllable. Moreover, many languages employ differences in relative prominence not only between syllables in a word, but also between different words in a phrase or utterance. In particular, there is often one dominantly prominent word within some larger prosodic domain, such as the intonational phrase (Wagner 2002). This word is typically marked by placing a pitch accent, that is, a local maximum in the F0 contour of the utterance, on the primary stressed syllable, but also by a durational increase of the word or parts thereof, as will be examined below. We shall employ the general term *accent* to refer to this notion of prominence. No more fine-grained differentiation on linguistic grounds will be attempted, as the model we are going to introduce will be rather simple and largely agnostic to such differentiations.

In the discussion of Aylett and Turk (2004) in the preceding chapter, we already hinted at the importance of prominence patterns for the structuring of information in speech. On this view, rendering a linguistic unit prominent relative to its context may be understood as a strategy employed by speakers to draw the listener's attention to this unit because it transmits particularly important information. This perspective is most straightforwardly demonstrated for accent: accentuation typically coincides with the part of the message that the speaker considers to be most important in the context in which the utterance is produced, often coinciding with *new information*. Thus, the utterance *Tim ate some BREAD yesterday* (with capitalization indicating accentuation of *bread*) would be expected as an answer to the explicit or implicit question what it was that Tim ate yesterday; that it was bread is the new information. *Bread* is therefore highlighted by accentuation against the rest of the sentence, which is given by the preceding discourse context.[2] By contrast, shifting the accent, or reallocating the *focus* (Bolinger 1958, Ladd 2008) of the utterance to *yesterday* would be expected in a context where it is already established that Tim ate bread and it has been asked explicitly or implicitly when this has happened. As for lexical stress, a similar argument may be deployed based on the cross-linguistic observation that stress tends to fall on the root

[2]The same accent pattern would be used if all the information in the sentence were new, a situation referred to as *broad focus* or *citation form* (Wagner 2002).

morpheme in morphologically complex words (Echols and Newport 1992). Perhaps more importantly, the existence of a more or less regular patterning of prosodically strong and weak elements in speech provides listeners with potential cues for word segmentation, even though they do not necessarily make use of this information (Cutler and McQueen 2014 and references therein). Some languages also employ stress placement in *minimal stress pairs*, i.e., to distinguish between segmentally identical lexemes such as *OBject* and *obJECT* in English.

This functional perspective on prominence lends itself well to interpretation within H&H theory. On this view, the greater prominence of some syllables and words in relation to their environment may be seen as a consequence of a greater demand for perceptual clarity, so as to ensure that these important items are successfully communicated. This has been explicitly argued by De Jong (1995) in his account of prominence as "localized hyperarticulation". Alluding to the notion of "sufficient contrast" in H&H theory, De Jong, hypothesizes that prominence should increase various types of phonemic contrasts, which is exactly what he finds in his articulatory study: stressed syllables involve not only longer, faster and greater jaw movements than unstressed ones, but also enhanced distinctions such as backness and roundedness in vowels. Other studies have reported some evidence supporting this claim for linguistically meaningful contrasts such as distinctive vowel quantity (De Jong 2004 for English, Heldner and Strangert 2001 for Swedish) or postvocalic voicing (De Jong and Zawaydeh 2002 for English, but not for Arabic). In the following review, we will encounter one more example of this kind, enhancement of lexical stress contrasts in accented environments.

3.2.2 Review

A large body of research shows that in many languages, segments in lexically stressed syllables are longer than in unstressed syllables, ceteris paribus, although the magnitude of stress-induced lengthening may vary cross-linguistically (e.g. Delattre 1966, Lehiste 1970, Prieto et al. 2012). De Jong (1995) calls for caution in making such comparisons in some languages, notably English, in which stressed and unstressed syllables tend to exhibit vowel quality differences, with unstressed vowels being typically more centralized. Yet, we would argue that this is a direct consequence of stress, rather than a confound. Moreover, studies featuring control for vowel quality differences document reliable durational effects of stress (e.g. Okobi 2006, van Santen 1992 for English). A further problem present in many older studies is that stress and accent have often not been properly disentangled – if the effect of stress is investigated in sentences of the *Say X again* type, as has often been done, the target word is likely to bear a nuclear accent,

hence the data do not reveal anything about stress effects in unaccented contexts. Studies that did account for these factors have provided evidence that stress is marked by increased duration even in unaccented words (e.g. Cambier-Langeveld and Turk 1999, Okobi 2006, van Santen 1992 for English, Sluijter and Van Heuven 1996 for Dutch, Dogil and Williams 1999 for German, Heldner and Strangert 2001 for Swedish).

In polysyllabic words, listeners tend to perceive *secondary stress*, that is, additional subordinate prominences besides the main stress in many languages. Results regarding durational consequences of secondary stress are mixed. Findings from reiterant speech studies, in which subjects were instructed to produce repetitive syllable sequences with stress patterns of existing words, suggest that secondary stress is marked only by some speakers in American English (Nakatani et al. 1981) and Indonesian (Adisasmito-Smith and Cohn 1996), whereas it seems to be consistently marked in Dutch (Rietveld et al. 2004). As for American English, evidence for secondary stress marking comes from the corpus analysis by van Santen (1992): controlling for vowel and postvocalic consonant identity as well as within-word position, he finds that speakers do lengthen secondary stressed relative to unstressed vowels in accented and unaccented words. Kleber and Klipphahn (2006)'s production study with stressed, secondary stressed and unstressed vowels from word-initial syllables with matching onset and coda consonants does not suggest durational consequences of secondary stress in German.[3] In a similar study on American English, Plag et al. (2011) find no durational differences between primary and secondary stressed English vowels in both accented and unaccented contexts. No comparison with unstressed vowels is attempted in this study, but since the lengthening of primary stressed relative to unstressed vowels in English is well-established, it may be conjectured from its result that secondary stressed vowels are also lengthened relative to unstressed ones.

Much research has been devoted to *stress shift*, the hypothetical tendency of speakers to change prominence patterns of words in order to avoid adjacency of stressed syllables, as in the case of *thirTEEN*, which supposedly becomes *THIRteen* in the phrase *THIRteen MEN* (capitalization indicates stress placement). However, evidence from English suggests that stress shift, if it occurs, seems to be a primarily perceptual phenomenon with very little acoustic consequences (e.g. Grabe and Warren 1995). Vogel et al. (1995) do find a subtle durational effect: *-teen* in the above example is slightly shorter in a stress shift context than in a context that does not trigger stress shift.

[3] A possible issue with this study is that in more than half of the test items, the secondary stressed vowel comes from a longer word than the unstressed vowel. As we will see later, vowels are shorter in longer words if they are accented, which is likely to be the case in Kleber and Klipphahn (2006)'s data, given their use of *Say X again*-type carrier sentences. Durations of secondary stressed vowels may thus be biased downward in this study.

The remainder of this review will be concerned with phenomena related to accent. As outlined above, we define accent as the dominant prominence of a particular word relative to other words within a larger prosodic domain. A large body of research indicates that words are longer when they are accented than when they are not (e.g. Cambier-Langeveld 2000 for Dutch and English, Heldner and Strangert 2001 for Swedish, Dogil and Williams 1999 for German, Botinis 1989 for Greek, Ortega-Llebaria and Prieto 2011 for Spanish and Catalan). Most of these studies have utilized controlled reading paradigms, triggering accentuation or deaccentuation of a target word by capitalizing either the word itself or other words in the surrounding carrier sentence in reading materials given to subjects. We deem it irrelevant for the present purpose whether accentual lengthening is a purely mechanical consequence of the need to accommodate the pitch movement typically found in accented words, or an independent effect (cf. White 2002 and references therein). Suffice it to note that accentual lengthening is a consequence of prominence, whether it is mediated by pitch movement or not.

So far, we have been referring to accent as word prominence, but mixed results have been reported as to which parts of an accented word actually undergo lengthening. Turk and Sawusch (1997) report lengthening of the stressed syllable and following, but not preceding unstressed syllables in accented words in English, whereas Turk and White (1999), Cambier-Langeveld and Turk (1999) and (White 2002) report lengthening of all syllables in accented words in English, and Cambier-Langeveld and Turk (1999) observe this also for Dutch. Heldner and Strangert (2001) reports that in Swedish, only the stressed and the immediately following unstressed syllable are lengthened in accented words. Complex patterns may arise in very long words: Suomi (2007) observes no significant accentual lengthening in final syllables of tetrasyllabic words with initial stress in Finnish, and for English, Dimitrova and Turk (2012) find that only some locations within tetrasyllabic words are lengthened. Word length may also be a factor in explaining results from similar studies on German: Samlowski et al. (2014) find reliable accentual lengthening only in the stressed syllable in tri- and tetrasyllabic words, and Dogil and Williams (1999) reports no accentual lengthening at all in five-syllable words in German.

Thus, the generalization may be proposed that accentual lengthening affects a multisyllabic domain delimited by word boundaries, but not necessarily involving all syllables in a word. The most important question for our review concerns the distribution of accentual lengthening within this domain. Turk and White (1999) find varying degrees of accentual lengthening in trisyllabic English words such as *property*: 23% in the (stressed) initial, 11% in the medial, and 14% in the final syllable of the word, compared to the unaccented baseline. This seems to suggest that position within the word, or lexical stress, or a combination of both, may mediate accentual lengthening in English; however, the authors rightly hesitate to make such claims, because the phonetic material is of course

not the same in the three syllables. This is also the case in most of the other studies reviewed above.

Comparing accentual lengthening in item pairs such as *bake enforce* and *bacon force*, Turk and Sawusch (1997) and Cambier-Langeveld and Turk (1999) find that lengthening is stronger and more consistent in word-final than in word-initial unstressed syllables in English and Dutch. Comparisons between stressed and unstressed syllables are, again, not possible because of segmental differences. The studies by Sluijter (1995) on English and Sluijter and Van Heuven (1996) on Dutch allow for directly comparing the effect of accent on stressed and unstressed syllable durations, by utilizing minimal stress pairs and words composed of reiterant syllables such as "baba" in controlled reading experiments. Generally, no significant interactions between stress and accent are found in these studies, but this conclusion is based on omnibus ANOVAs on the combined datasets. Once within-word position is controlled, a complex picture emerges (cf. Cambier-Langeveld and Turk 1999): in word-initial position, lengthening percentages for stressed syllables are between two and three times as large as for unstressed syllables. In contrast, the proportional effect of accent in word-final position is roughly the same in stressed and in unstressed syllables. Referring to earlier proposals (Klatt 1976, Nooteboom 1972), Sluijter and Van Heuven (1996) suggest that prominence-induced lengthening becomes weaker in the presence of final lengthening in order to prevent the syllable from exceeding a hypothetical upper duration boundary.

This explanation would suggest that accentual lengthening interacts with stress and, additionally, with word-final lengthening: stressed syllables are lengthened more strongly in accented words than unstressed syllables, but this difference diminishes in word-final position. However, an alternative explanation is conceivable: since accentual lengthening is approximately similar percentage-wise in word-initial stressed and word-final unstressed syllables, results reported so far are also consistent with the hypothesis that accentual lengthening just spreads rightward from the stressed syllable onset, affecting all syllables equally regardless of their stress status, whereas there is only some residual, "spill-over" lengthening to the left of the stressed syllable onset. Turk and White (1999)'s finding of approximately twice as much accentual lengthening in a word-initial stressed than in following unstressed syllables in English argues for inherently stronger accentual lengthening in stressed than in unstressed syllables, as do Heldner and Strangert (2001)'s Swedish data and findings on Romanian by (Manolescu et al. 2009); yet, in these studies, phonetic materials are not identical in stressed and unstressed syllables.

The corpus study by van Santen (1992) supports the view that accentual lengthening affects stressed vowels inherently more than unstressed ones. van Santen (1992) analyzes vowel durations from utterance-medial words produced by two speakers, controlling for

vowel identity, post-vocalic consonant and within-word-position. The central finding is that once these factors are controlled, stressed vowels are on average 39% longer in accented words, but only 22% longer in unaccented words than unstressed vowels. which suggests . From these figures it can be calculated that accentual lengthening is some 15% stronger in stressed than in unstressed vowels. Similar results have also been found in an experimental study of bisyllabic words consisting of reiterant syllables in Greek (Botinis 1989).

These studies thus do not suggest that accentual lengthening in stressed and unstressed syllables is similar in word-final position. A possible issue with van Santen's analysis is that he does not report cell frequencies. In particular, since the majority of lexical words in English have initial stress (Cutler and Carter 1987), one might suspect that there are not many observations for polysyllabic words with stress on the final syllable in an unbalanced database such as the one utilized by van Santen. As samples sizes are even further reduced by the author's data partitioning techniques, it is possible that van Santen's failure to find an interaction between stress, accent and word-final lengthening is due to lack of statistical power. In any case, this explanation does not apply to the study by Botinis (1989), which is similar in design to those by Sluijter (1995) and Sluijter and Van Heuven (1996). Given this state of matters, one might tentatively conclude that there is a cross-linguistic tendency for accentual lengthening to be stronger in stressed than in unstressed syllables, with word-final position as a possible complicating factor in some languages.

One language for which this relationship does not seem to hold, however, is Spanish. Ortega-Llebaria and Prieto (2011) investigate durational effects of stress and accent in Spanish and Catalan, using minimal stress pairs in these languages. The authors do not report interactions between both effects explicitly, but graphical presentation of results suggests that whereas Catalan follows the pattern reported for other languages, Spanish does not. Data presented in Kim (2011) also suggest a simple additive combination of the lengthening effects of stress and accent in Spanish. The difference between Spanish on the one and Catalan and English on the other hand would appear to suggest that vowel reduction in unstressed syllables may be a relevant factor. Indeed, spectral reduction of unstressed vowels has also been reported for Greek (Fourakis et al. 1999), which patterns with English and Catalan regarding the stress-accent interaction. It would be an interesting perspective to investigate this possible connection in a larger cross-linguistic study.[4]

[4]Okobi (2006) reports on a study of stress correlates in English that controls for pitch accent and, additionally, explicitly examines full vowels in unstressed syllables. Unfortunately, the author reports durational results only in terms of differences between vowels within the same target word, which makes it impossible to assess possible stress-accent interactions.

Another factor that has been found to influence the distribution of accentual lengthening is the length of the accented word, in terms of the number of syllables. For example, Turk and White (1999) find that accentual lengthening of the primary stressed syllable is larger in both absolute and proportional terms in monosyllabic (e.g. "*BAKE* enforce" vs. "*bake* ENFORCE") than in bisyllabic target items "*BAC*on force" vs. "*bac*on FORCE"). Total syllable duration is also greater in the monosyllabic than in the bisyllabic condition. The same pattern of results is reported by White (2002): all constituents of the primary stressed syllable are lengthened less strongly by pitch accent if the word contains more syllables. The same applies to the duration of unstressed syllables immediately adjacent to the stressed syllable in the bi- versus trisyllabic condition, such as /-ən/ in *mason* versus *masonry* and /kə-/ in *commend* versus *recommend*. These findings suggest inverse relationships between word length and durations of syllables in accented words. Such inverse relationships have been taken to support broader claims about the temporal organization of speech, which will be discussed at length in a subsequent section of this chapter. For the moment, we may note that the relationship between the strength of accentual lengthening and word length may offer an explanation for the negative findings on accentual lengthening in very long words discussed above.

3.2.3 Summary

We have seen that prominence is robustly signaled by timing contrasts in many languages, at the word (stress) and at the phrase or utterance level (accent). As for accent, the temporal lengthening effect is distributed over some larger domain, such as the word. This lengthening is not uniformly distributed: It affects stressed more than unstressed syllables, although differences may diminish in word-final position, possibly due to an interaction with word-final lengthening. Moreover, the length of the domain seems to play a role: accentual lengthening of individual syllables, as well as total lengthening of the whole word diminishes in longer words, as will be discussed in more detail below.

Prominence is a particularly interesting phenomenon for this work, because its interpretation as "localized hyperarticulation" suggests a straightforward implementation in a H&H-based optimization account: prosodic prominence can be simulated in such an approach by means of weighting factors that locally boost perceptual clarity. Whereas the lengthening effects themselves will be obvious consequences of such a strategy, it will be most interesting to see whether the stress-accent interaction described above will fall out automatically from the modeling assumptions. The same goes for other interactions involving prominence.

3.3 Positional Effects

3.3.1 Introduction

The timing of suprasegmental units in speech is influenced not only by prominence, but also by their position relative to other units. A large body of research has demonstrated that *boundary-adjacent lengthening* is a pervasive phenomenon in speech: segments are longer if they appear close to the boundaries of certain larger constituents than when they appear constituent-internally. In particular, final lengthening at various levels of the prosodic hierarchy is well-documented. In this section, we shall review research on the location and scope of these effects, as well as potential interactions with prominence.

Positional effects on speech timing have traditionally been explained as automatic consequences of biomechanical properties of the human vocal tract. Lengthening at the end of prosodic constituents in particular was hypothesized to stem from inertial properties of the vocal tract muscles (e.g. Fowler 1990). On this view, final lengthening is interpreted as an instance of a hypothesized general tendency of motor systems to decelerate towards the end of movement trajectories. Indeed, Edwards et al. (1991) observe that the articulatory changes involved in final lengthening resemble those involved in the reduction of overall speaking rate.

Proponents of a different position claim that final lengthening, like prominence, has a primarily perceptual motivation, signaling upcoming boundaries to the listener. This position is put forward by White (2014), who argues that final lengthening is actively employed for linguistic purposes, although he does not rule out that it originates from biomechanical properties of the speech organs. He does, however, refer to findings showing that final lengthening actually has to be learned in the course of first language acquisition (Snow 1994). Referring to this work, White brings up the possibility that final lengthening "is therefore an acquired skill rather than a product of articulatory constraints" (2014:39).

In a similar vein, Turk and Shattuck-Hufnagel (2014a) put forward the hypothesis that boundary marking in speech may be related to language redundancy, comparable to Aylett and Turk (2004)'s ideas about prominence structure discussed in Chapter 2: they argue that dividing the speech stream into shorter phrases reduces uncertainty by limiting the number of possible word boundary parses. On this view, constituent-final lengthening would be interpreted as means of information packaging for the listener. Finally, the "biomechanical" explanation draws much of its appeal from the idea that final lengthening effects in speech are a specific instance of a general pattern in the human motor system, but the generality of such a pattern may, in fact, be debatable: for

example, Turk and Shattuck-Hufnagel (2015) find no evidence for series-final lengthening in a finger wagging experiment. In view of this body of evidence, we consider the "communicative" explanation of positional effects in speech to be more likely than the "biomechanical" explanation, although no definitive verdict can be given at this point, and as stated by White (2014), both accounts are also not necessarily mutually exclusive.

3.3.2 Review

One of the best-documented findings in speech timing research across various languages is lengthening at the end of large prosodic constituents, such as phrases or utterances Fletcher (2010). Klatt (1976) in his review of speech timing effects in English reports that a vowel in a phrase-final syllable may be twice as long as the same vowel in non-final position. This lengthening effect occurs before silent pauses, but also if the phrase boundary is not followed by a pause. An important question that arises is which kind of phrase triggers the effect. First, an utterance may be divided into *prosodic* and *syntactic* phrases, which often but not necessarily correspond to each other; as for this question, White (2002) and research reviewed therein suggests that it is the prosodic phrasing that mediates syntactic structure and effectively triggers suprasegmental timing effects.

The question remains what type of prosodic phrase triggers final lengthening effects. Several researchers have posited *prosodic hierarchies*, systems of successively larger prosodic constituents nested within each other. Distinctions are made, for example, between the *phonological phrase*, *intonational phrase* and the *phonological phrase* (Nespor and Vogel 2007), or between *minor* and *major intonational phrases* (Selkirk 1986). We will not dwell on these distinctions, because the model we are going to introduce below is agnostic towards the phonological features that characterize different levels of prosodic hierarchies. We simply note that prosodic phrases can be conceived as hierarchically organized, which seems to be reflected by timing characteristics: stronger lengthening effects are observed at higher-level boundaries (e.g. Wightman et al. 1992 for English, Horne et al. 1995 for Swedish, Prieto et al. 2012 for English, Spanish and Catalan).

Whereas lengthening towards the end of phrasal units in speech is a probably universal phenomenon, it is less clear if there is reliable word-final lengthening in the absence of phrasal boundaries. The problem, as hinted at by Klatt (1976), is precisely that word and higher-level boundaries are difficult to separate. Klatt (1976) reports mixed results from earlier research for word-final lengthening in the absence of phrasal boundaries in English and states that where lengthening effects are reported, they may be too small to be perceptually relevant. The reiterant speech study by Nakatani et al. (1981) yields evidence for word-final lengthening in American English even in the absence of phrasal

boundaries. Caution may be warranted in drawing conclusions from this finding, however; it is not unlikely that speakers employ an exaggerated style of contrastive prosodic marking in reiterant productions, and the authors also do not supply information on the distribution of pitch accents in their data. Beckman and Edwards (1990) analyze productions of non-phrase-final *pop opposed* versus *poppa posed* sequences produced by five speakers of American English at different rates and find word-final lengthening only in the slow rate condition and for some speakers.

The corpus analysis by van Santen (1992) does support word-final lengthening in the absence of at least utterance boundaries in American English. As we have seen in the previous section, Sluijter and Van Heuven (1996) invoke word-final lengthening as a possible cause for differences in accentual lengthening in word-final syllables in Dutch and English. A study on utterance-medial and likely nuclear-accented reiterant CVCV-words produced by five speakers of Greek (Arvaniti 2000) provides evidence for word-final lengthening of stressed, but not of unstressed vowels. However, no clear effect is discernible for syllable durations, due to mixed durational behavior of the onset consonant /p/, depending on the vowel. The similar study by Botinis (1989), moreover, yields no evidence whatsoever for word-final lengthening in stressed or unstressed contexts, regardless of accentuation. Prieto et al. (2010) study identical syllables in word-final and word-penultimate position in accented words produced in sentence contexts by speakers of Spanish and Catalan and find no consistent evidence for word-final lengthening. In a methodologically comparable study on Italian, d'Imperio and Rosenthall (1999), report that at least in open syllables, vowels are actually *shorter* word-finally than elsewhere.[5] This small survey suggests that word-final lengthening in the absence of higher-level boundaries may occur in some languages, where it may be style- or speaker-specific, but it not universally attested.

There is ample evidence for effects of *initial* position in certain prosodic constituents on acoustic and articulatory characteristics of speech. These *initial strengthening* effects, however, are quite different from effects at final boundaries of prosodic constituents, as they are narrowly localized on word-initial segments, and do not usually trigger durational effects at the scale of domain-final lengthening (see Cho et al. 2007 for an overview). Moreover, initial strengthening does not necessarily surface in acoustic durations. For example, Cho and Keating (2009) find clear effects of utterance-initial position on electropalatography measures, but no effects on the acoustic durations of word-initial nasals in American English, whereas different degrees of prominence are clearly mirrored in acoustic durations. Although it qualifies as a suprasegmental effect, we will therefore not consider initial strengthening in this work, assuming that this is a

[5] As for stressed vowels, van Santen and d'Imperio (1999) note that this may have to do with segmentation criteria, as stressed vowels in word-final position are often heavily glottalized in Italian.

task for fine-grained articulatory modeling. The remainder of our review of positional effects in speech will be concerned with final lengthening at major prosodic boundaries, without attempting to introduce fine-grained distinctions between boundary types.

Evidence from a number of languages suggests that final lengthening before major prosodic boundaries is not limited to the syllable that is directly adjacent to the boundary: stressed syllables undergo lengthening in utterance-penultimate or antepenultimate position (e.g. van Santen 1992, White 2002 for English, Cambier-Langeveld 2000 for English and Dutch, Kohler 1983 for German, Nakai et al. 2012 for Finnish, Berkovits 1994 for Hebrew). Turk and Shattuck-Hufnagel (2007) observe that in American English, even stressed vowels in tetrasyllabic words with initial stress undergo some lengthening if the word occurs utterance-finally, whereas word-medial unstressed syllables in utterance-final words do no undergo reliable lengthening. Results on final lengthening in syllable positions not directly adjacent to the boundary may also be mediated by pitch accent: for example, Cambier-Langeveld (2000) for English and Nakai et al. (2012) for Finnish observe lengthening in utterance-penultimate stressed syllables only in unaccented contexts. Berkovits observes a small lengthening effect (9%) also in word-initial unstressed syllables if a word with final stress occurs in utterance-final position, but it is not clear if this difference reaches statistical significance. What the cited studies generally agree on is that the magnitude of final lengthening decreases with distance to the boundary: for example, van Santen (1992) reports between 65 and 93% lengthening of stressed vowels in utterance-final syllables relative to comparable utterance-medial vowels, compared to 25% lengthening in utterance-penultimate syllables. It is not known if utterance-penultimate or -antepenultimate vowels are lengthened as well if there is yet a word boundary between the target vowel and the end of the utterance.

The foregoing review has already hinted at possible interactions between final lengthening and lengthening due to prominence. As for lexical stress, surprisingly few studies have investigated this question in a controlled fashion. In his corpus analysis of American English by van Santen (1992) compares utterance-final lengthening in stressed and unstressed vowels, controlling for the identity of the vowel and the postvocalic consonant. He finds approximately similar lengthening ratios for both categories in utterance-final versus medial position. One caveat is that this conclusion is based on 70 observations from two speakers. Moreover, all observations come from pitch-accented words, so that inference about unaccented contexts is not possible. Nakatani et al. (1981) have investigated the interaction of stress and final lengthening in American English in an experimental study on reiterant "ma" syllables, instructing speaker to produce them so as to match the stress patterns of existing words. No statistical analyses are reported, but graphical presentation of results suggest that the absolute amount of lengthening due to lexical stress is roughly identical in phrase-medial and phrase-final position, which is

compatible with van Santen (1992)'s data. No information on the distribution of pitch accents is given. By contrast, Klatt (1976:1214) states that the stressed/unstressed durational contrast in English is "largest in a phrase-final syllable", but it is not clear what data this statement is based on.

Campos-Astorkiza (2014) reports an experimental comparison of stressed and unstressed vowel durations in word-final open syllables in phrase-medial versus phrase-final position, produced by four speakers of Tuscan Italian. Her results suggest that unlike in American English, the durational contrast between stressed and unstressed vowels is proportionally greater in phrase-final than in phrase-medial position in Tuscan Italian, which is also robustly observed for all four speakers. One caveat is that pre-vocalic consonantal context is not completely uniform across stress conditions, but confounding effects due to this source are probably small and unsystematic. Target vowels in this study are highly likely to bear a nuclear pitch accent.[6] The study on bisyllabic Hebrew words with initial and final stress in utterance-medial and final position by Berkovits (1994) also provides evidence bearing on the interaction between stress and final lengthening: the nucleus of the word-final syllable is /ɪ/ in both stress conditions, and since the postvocalic consonant is always a voiceless stop, results for utterance-final lengthening in word-final stressed and unstressed syllables are comparable. These data suggest that proportional final lengthening in Hebrew, at least for this vowel, is greater in unstressed than in stressed syllables, 57 vs. 38%.

The interaction between lengthening induced by boundary adjacency and accent, i.e., prominence at the phrasal level, has been investigated in more studies. Most of these studies report duration measurements for entire words. For example, Cooper et al. (1985) investigate the effect of "contrastive stress", i.e., contrastive focal accent on word duration in different utterance positions in American English. They make no effort whatsoever to control for segmental differences or even syllable count in test words in different positions, but since their study includes a variety of different materials, one may assume that segmental effects should be more or less randomly distributed with regard to utterance position. Their results suggest that contrastive focus lengthens words in utterance-final position by roughly 17% on average, compared to approximately 40% lengthening on average in other positions.

The above-mentioned study by Cambier-Langeveld (2000) documents an interesting difference between English and Dutch regarding the interaction of final and accentual lengthening: in English, absolute final lengthening of word duration is equal in accented and unaccented contexts. In Dutch, there is evidence for an interaction: accented and unaccented durations converge in utterance-final position, i.e., there is no accentual

[6] Valentina Schettino is gratefully acknowledged for this native intuition.

lengthening at all in utterance-final words. While English and Dutch differ in that *absolute* accentual and final lengthening interact in the latter but not in the former language, both languages are similar in that *proportional* accentual lengthening is stronger in non-final than in final contexts, even though the difference is more pronounced in Dutch. As for English, this conclusion is also supported by comparisons of accented and unaccented materials in utterance-medial and -final position reported in White (2002).

Heldner and Strangert (2001) examine durational effects of focal accent in different positions in Swedish utterances. Their results suggest that in contrast to English and Dutch, proportional accentual lengthening of bisyllabic words in Swedish tends to be somewhat stronger in utterance-final than in initial and medial position. Differences are small, however, and reach significance only for one of two test words. No results for individual syllables are reported. Results more in line with English and Dutch are reported in Roosman (2006) for Betawi Malay and Toba Batak, two Austronesian languages: using a methodology similar to that of Cambier-Langeveld (2000), the author finds that final lengthening of bisyllabic words is slightly less pronounced if the target words are accented. Complete absence of lengthening due to focus in utterance-final position is reported for both real and reiterant words in Italian by Farnetani and Zmarich (1997). Nakai et al. (2012)' Finnish data suggest that proportional accentual lengthening of stressed vowels in Finnish is somewhat stronger in utterance-medial than in final position, whereas there are no differences between both positions for unstressed vowels. There are, however, large individual differences between test items attributable to the contrastive vowel quantity in Finnish.

3.3.3 Summary

Lengthening at major prosodic boundaries is a pervasive and probably universal phenomenon in speech. Some cross-linguistic tendencies in the distribution of final lengthening can be pointed out: it starts earlier than the syllable that is directly adjacent to the boundary, possibly at the main stress of the last word in the utterance. Within this domain, lengthening is progressively stronger towards the boundary, although there is the possibility that unstressed syllables between the last prominence in the prosodic domain and its boundary are "skipped" by final lengthening (Dimitrova and Turk 2012). Our review suggests that in most languages, prominence and positional effects do not enhance each other in the way different levels of prominence do: prominence effects on duration tend to be proportionally weaker in boundary-adjacent than in constituent-medial position, sometimes to the extent that durational contrasts due to prominence are completely neutralized in constituent-final position. Thus, while we favor the hypothesis that positional effects on suprasegmental speech timing are actively employed

for communicative purposes in a fashion similar to prominence effects, the behavior of both classes of effects is to be quite different, rendering it less obvious how positional effects should be incorporated in the envisaged model of speech timing.

3.4 Constituent Length Effects

3.4.1 Introduction

In addition to effects of prominence and position, it has often been claimed that durations of syllables or parts thereof are influenced by the number of syllables included in larger prosodic constituents. It is usually assumed that these relationships are of a compensatory nature, such that syllables or vowels become shorter as the number of them in a given larger constituent increases. We will refer to these effects as *constituent length effects*.[7] Constituent length effects are a natural prediction of theories which assume that speakers attempt at regularizing the durations of certain prosodic constituents, in particular the widely discussed *isochrony hypothesis* (Abercrombie 1967, Pike 1945). As we shall see later, a number of explanatory accounts of constituent length effects have been proposed; in particular, the empirical status of such effects is of crucial importance for the currently most prominent class of explanatory computational models of suprasegmental speech timing, coupled-oscillator models. On the other hand, the existence of the entire class of constituent length effects as an independent phenomenon has been called into question in recent work (White 2002, 2014). In what follows, we will provide a concise review of current research.

3.4.2 Review

The study by Port (1981) provides a classical, although by no means the earliest example of investigations of constituent length effects and shall serve us for expository purposes. Port reports measurements of stressed vowel duration in sentences such as *I say d[iː]p/d[iː]per/d[iː]perly every Monday.* The sentences thus differ in the number of unstressed syllables following the test syllable within the same word. Port reports the duration of the vowel in the test syllable to vary inversely with the number of following syllables in the word: the stressed vowel is shorter in bisyllabic (*deeper*) than in monosyllabic words (*deep*), and shortest in trisyllabic words (*deeperly*). This inverse relationship is asymptotic, i.e., the difference in stressed vowel duration is greater between monosyllables and bisyllables than between bisyllables and trisyllables. Port reports

[7] In preferring this term over the widely-used term *polysyllabic shortening*, we follow White (2002), who reserves the term polysyllabic shortening for constituent length effects at the word level.

similar effects, though weaker in magnitude, also for the preceding and following stop closure, indicating that the syllable as a whole is affected.

Similar studies using reiterant materials (Lindblom and Rapp 1975 for Swedish, Nooteboom 1972 for Dutch) also allow for assessing the effect in unstressed syllables, where it is less pronounced than in stressed syllables. Constituent length effects have been investigated in many studies using methodologies similar to the above ones (e.g. Klatt 1973, Lehiste 1972 for English, Lindblom 1968, Strangert 1985 for Swedish, Braun and Geiselmann 2011, Farnetani and Kori 1986, Vayra et al. 1999 for Italian, Kohler 1983, Rietveld 1975 for German). These studies report similar asymptotic shortening of stressed vowels or syllables as a function of the number of syllables in larger constituents.

However, White (2002), in his extensive review of earlier work on constituent length effects, points out two major methodological shortcomings present in most previous studies on the subject: first, most previous results are ambiguous on the question which constituent is actually responsible for the shortening effect. For example, the effect observed by Port (1981) could also be a consequence of the increasing syllable count in the *utterance* rather than the target word, as it has been conjectured that speakers talk faster in longer utterances (e.g. Klatt 1976, Lehiste 1972). Moreover, the above-mentioned isochrony hypothesis posits that speakers of "stress-timed" languages such as English attempt at placing stressed syllable onsets at temporally regular intervals. This would predict that the shortening is triggered by the syllable count in the *inter-stress interval* (ISI), here presumably ranging from the target syllable to the first syllable of *again*, and thus largely overlapping with the target word. It is also not clear if adding syllables *before* the target vowel in the target word would trigger any shortening. White (2002)'s second and more important concern is that many previous investigations of constituent length effects have failed to control for phrasal prominence: White notes that the word containing the target syllable in Port (1981)'s study, as well as in most of the other studies listed above, is likely to bear a nuclear accent. As we have seen, accent seems to have durational consequences throughout the word, so that it is very likely to interfere with alleged constituent length effects.

The study by Turk and Shattuck-Hufnagel (2000) avoids many of the design flaws present in earlier studies of constituent length effects. The authors report vowel duration measurements from ambiguous phonetic sequences such as *tuna choir/tune a choir/tune acquire* produced utterance-medially in uniform carrier phrases by six speakers of American English. Thus, they sidestep the need to introduce additional syllables, which would increase syllable count in both the target constituent and the whole carrier sentence. Turk and Shattuck-Hufnagel (2000) use capitalization of certain words in their experimental reading materials to control accentuation of target items. This design enables them

to manipulate word length exclusively and to assess possible interactions with phrasal prominence. Turk and Shattuck-Hufnagel find that both the first and the last vowel in the test sequence are longer in a monosyllabic than in a bisyllabic word – that is, /uː/ is longer and /aɪ/ is shorter in *tune acquire* than in *tuna choir*. Differences are greater in accented contexts and if the stressed syllable is followed rather than preceded by an unstressed syllable within the same word. Effects are generally rather small, amounting to differences of roughly 10% in accented and less than 5% in unaccented environments. These results suggests shortening as a function of syllable count in the word as a unified explanation, with the qualification that adding syllables after the stressed syllable has a stronger effect than adding syllables before it.

The study by White (2002), which we already briefly discussed in the preceding section, is, in our opinion, the most carefully-designed investigation of constituent length effects to date. It consists of reading materials such as the following (cf. White 2002:148; target syllables are printed in italics; "left-headed" and "right-headed" denote words with initial and final stress, respectively):

Left-headed +Accent:	**Left-headed -Accent**:
I saw the *MACE* unreclaimed again	I SAW the *mace* unreclaimed AGAIN
I saw the *MAS*ON reclaimed it all	I SAW the *mas*on reclaimed it ALL
I saw the *MAS*ONRY cleaned again	I SAW the *mas*onry cleaned AGAIN

Right-headed +Accent:	**Right-headed -Accent**:
John saw Jessica *MEND* it again	JOHN saw Jessica *mend* it AGAIN
John saw Jessie COM*MEND* it again	JOHN saw Jessie com*mend* it AGAIN
John saw Jess RECOM*MEND* it again	JOHN saw Jess recom*mend* it AGAIN

This experimental design allows for studying constituent length effects at the word level beyond the mono-versus bisyllabic comparison. It also allows for examining effects on the duration of the unstressed syllable added next to the stressed syllable between the bi- and the trisyllabic condition, since its segmental makeup and context are kept constant. Utterance length is kept constant by removing a syllable from another word in the sentence for every syllable that is added to the target word. The length of the ISI is also the same across conditions, presumably ranging from the test syllable to "-claimed" or "-gain", respectively in the above examples. Target words are in phrase-medial position, so that there should be no interference from boundary-adjacent lengthening. White finds that adding syllables either before or after the stressed syllable consistently shortens stressed syllable duration if the word is pitch-accented. In the unaccented condition,

word length only has a significant effect in left-headed words, and even there, it is minute in absolute and proportional terms.

Based on these results, White (2002) argues that the effect in accented words is not a constituent length effect in the sense of a tendency towards delimiting or equalizing word durations, but rather follows from redistribution of accentual lengthening: "Because total lengthening is no greater in polysyllables than in monosyllables, the effect on particular subconstituents is attenuated when the word contains more syllables" (White 2002:3). White (2002) prefers to analyze the weak shortening effect of syllable count in unaccented left-headed words as progressive word-final lengthening, although he does mention the alternative possibility that the pattern indicates a genuine constituent length effect in the "word rhyme", the interval between the stressed syllable onset and the right word boundary. He concludes by suggesting a model of speech timing that consists exclusively of localized lengthening effects at the heads and edges of prosodic domains (stress, accent, final and initial lengthening at various constituent boundaries). The status of genuine constituent length effects in such a model would be "at best, marginal" (White and Turk 2010:469).

Studies on other languages with similar control of durational factors such as accent support White (2002)'s conclusion. Suomi (2007) reports duration measurements on vowels from mono- to tetrasyllabic accented and unaccented words in utterance-medial position, produced by six speakers of Finnish. He observes large differences between vowel duration in mono-versus polysyllabic words in some cases, especially in the accented condition (which would be explicable as a word-final lengthening effect), but very little evidence for differences in the direction of a constituent length effect beyond this comparison. Siddins et al. (2013) replicate White (2002)'s second experiment with German materials and six speakers. They report the same pattern of results, with the exception that in German, there seems to be no evidence whatsoever for a constituent length effect in unaccented words. It may be noted that an earlier study on German by Kohler (1983) does provide some evidence for shortening of stressed vowels by following unstressed syllables within the word and also across the right word boundary, but as this study is based on data from a single speaker – the author himself – caution may be warranted in interpreting its result.

White also addresses the question of constituent length effects at the utterance level, by including a series of test sentences with increasing syllable count while keeping the target word constant. It has sometimes been hypothesized that speakers talk faster in longer utterances (e.g. Klatt 1976, Lehiste 1972), but White finds no evidence for this claim and concludes that alleged constituent length effects at the utterance level are likely to be spurious. This conclusion is in line with results from an earlier experimental

study by Flege and Brown Jr (1982) on utterance-final productions of words composed of reiterant syllables in utterances of different length by eight speakers of American English. No effects of utterance length on stressed or unstressed vowel durations are reported in this investigation. The study by Hakokari et al. (2008) on two Finnish speech corpora complements this picture: the authors find weak but significant negative correlations between utterance length and segmental durations, which, however, vanish once utterance-initial and -final segments are removed from the analysis. The most likely explanation is that the apparent correlation between utterance length and segmental duration is a trivial consequence of the fact that utterance-initial and final phones, which are subject to boundary-adjacent lengthening, make up for a larger proportion of the total number of phones in short than in long utterances. On the other hand, van Santen (1992) does observe a correlation between utterance length and speaking rate whilst controlling for final lengthening. Yet, there may be other confounding factors involved in this finding; for example Crystal and House (1990) in a similar analysis report that a large portion of the durational variance attributed to utterance length is actually accounted for by the proportion of stressed syllables in the investigated utterances, and since van Santen (1992) does not report control for prominence in this analysis, it is possible that a greater proportion of stressed observations in shorter utterances may be responsible for his finding.

Some experimental results (Fowler 1977, Huggins 1975, Lehiste 1972, Van Lancker et al. 1988) suggest that stressed syllables in English are shortened by added unstressed syllables across the following word boundary, which is not investigated in the Turk and Shattuck-Hufnagel (2000) and White (2002) studies. For example, Fowler (1977) finds *fact* to be longer in *the FACT STARted the argument* than in *the FACT has STARted the argument* (stressed syllables in capitals). Such findings are compatible with a constituent length effect at the ISI level; however, White (2002) in his review notes that these studies do not provide evidence for any effect beyond the comparison between mono- and bisyllabic ISI, and reanalyzes it as "stress-adjacent lengthening": word-final stressed syllables are longer when followed by a stressed than when followed by an unstressed syllable. Shattuck-Hufnagel and Turk (2011) observe some evidence for an effect beyond this comparison in an analysis of stressed vowel duration in limerick stanzas produced by speakers of English. However, the effect varies across speakers and contexts, and the ecological validity of results from poetic speech, where speakers may put special emphasis on regularity, is arguably limited. Pamies Bertrán (1999) reports experimental results on potential influences of the number of following unstressed syllables on word-final stressed vowel duration from Spanish, Catalan, Portuguese, English, French, Italian, and Russian. The study features controlled materials (even though phonological

environments are not always held entirely constant) and several speakers for each language. Pamies Bertrán finds no evidence whatsoever for shortening of stressed vowels by the number of following unstressed syllables across the word boundary in any of the investigated languages.

In his corpus study of American English, van Santen (1992) investigates effects of four constituent-length variables on vowel duration: the number of following as well as preceding syllables in the word and in the ISI. His analyses are restricted to accented utterance-medial words, controlling for vowel identity and phonological environment. The values of the respective other three experimental variables are also held constant in each of the four analyses. Results of this controlled analysis suggest that only one of the experimental variables, the number of following syllables in the word, exerts shortening on vowels. van Santen (1992) adds the revealing comment that syllable count in the ISI does seem to exert a shortening effect on vowel duration if the location of a vowel relative to the following word boundary is not controlled. This result casts doubt on positive findings for a constituent length effects at the ISI level in English reported in other corpus studies (Bouzon and Hirst 2004, Campbell 1988, Kim 2006, Krivokapić 2013, Williams and Hiller 1994), who have not controlled their data with equal rigor.

As we said earlier, a general caveat about van Santen (1992)'s study is that his findings are restricted to data from two speakers, and to vowels from accented words, and a possible downside of his very careful data partitioning technique is that it presumably leads to rather small cell sizes, which may lead to issues with statistical power. Moreover, the question remains how to interpret the shortening effect of the number of following syllables in the word on vowel duration. In the review of similar findings by White (2002), we already hinted at the alternative possibilities of "word-rhyme compression" and progressive word-final lengthening. Neither van Santen (1992)'s nor White (2002)'s study offers a definitive answer to this question.

3.4.3 Summary

Constituent length effects in speech are widely supported, but their interpretation is not straightforward. What most studies do agree on is that such effects are reliably observed in words that bear a pitch accent. In unaccented contexts, there is less unambiguous evidence for such effects, and when they are observed, they are very weak in magnitude. The most in-depth study of such effects, White (2002), argues that constituent length effects are an artifact of phrasal prominence, but White's results do not completely rule out the possibility that such effects exist independently of prominence at the phrasal level. As we hinted at earlier and will investigate in more detail in Chapter 4 of this work,

various explanatory accounts of constituent length effects in speech have been posited. In a later part of this work, we will therefore report on an empirical investigation of such effects, which aims at answering open questions in this regard that have been posed by previous studies.

3.5 Effects of Overall Speaking Rate

3.5.1 Introduction

In this section, we will discuss effects of external conditions on suprasegmental timing patterns in speech. Speakers may change overall speaking rate, for example when being under time constraints, and such changes are reliably reproduced under experimental conditions (Xu 2010). Likewise, global variation on the H&H scale has been shown to influence global speaking rate: one of the most robust findings from studies that elicit "clear", listener-oriented (and thus, by hypothesis, hyperarticulated) speech is reduction of speaking rate in clear compared to normal conditions (e.g. Baker and Bradlow 2009, Picheny et al. 1986, Smiljanić and Bradlow 2005). The same has been found of *Lombard* speech, i.e., speech produced in noise (Šimko et al. 2014a and references therein), and for child-directed speech (Gallaway and Richards 1994). In any case, variations in speaking rate due to different causes may not be uniformly distributed throughout the speech signal, but interact with other timing effects. Consequently, we will consider studies of prominence and positional effects that examine speaking rate changes due to (experimentally induced) time pressure or global H&H variation an additional variable.

3.5.2 Review

Some controlled phonetic studies have investigated the durational interaction between prosodic prominence and speaking rate, by having subjects read sentences including target items that differ only in prominence level at different rates. Results are generally in agreement, with one notable outlier: Fourakis (1991) for American English, Fourakis et al. (1999) for Greek, den Os (1988) for Dutch (but not necessarily for Italian, where evidence seems to be inconclusive) and Nadeu (2014) for Spanish and Catalan find that stressed vowels shorten proportionally more than unstressed vowels in fast speech. By contrast, one experimental study on Dutch, Janse et al. (2003), reports stronger shortening of unstressed than stressed vowels and syllables in fast speech.

One methodological difference is that Janse et al. (2003)'s study alone differentiates stress and accent. However, as they report qualitatively identical results for the stressed-unstressed comparison in accented and unaccented contexts, this difference is not a likely cause for the different outcomes. Two methodological problems present in the Janse et al. (2003) study may be more severe: first, while Janse et al. (2003) balance stress position in the word and phonological vowel length, neither target vowel quality nor phonological environments are matched across stress conditions. By contrast, the other studies mentioned generally control for vowel identity and phonological environment by utilizing minimal stress pairs or words composed of reiterant syllables. Second, Janse et al. (2003) report assigning a default duration of 5 ms in cases where vowel duration was found to be too short to be reliably measured in their fast condition. They state that this was done in order to facilitate computation of fast/slow ratios for these vowels, and they explicitly criticize den Os (1988) for disregarding unstressed vowel tokens that were too short to be measured in their fast condition. If applied frequently, however, this method may create markedly bimodal duration distributions, rendering statistical analysis unreliable. Moreover, it may be hypothesized that at some point, vowels are actually deleted in fast speech, so that discarding non-measurable tokens would be justified.

For this to explain the difference between Janse et al. (2003)'s result and those from other studies, one would have to assume that unstressed vowels shorten less strongly, but are deleted earlier than stressed vowels. This may seem paradoxical, but it resonates with the well-established intuition of *incompressibility* in speech timing, i.e., the idea that there are lower thresholds to the durations of speech sounds for articulatory and perceptual reasons (Klatt 1973). Unstressed vowels may be hypothesized to be already closer to their compressibility threshold in slow speech than stressed vowels. This is precisely the explanation Fourakis (1991) and Fourakis et al. (1999) suggest for their finding of greater rate sensitivity of stressed than unstressed vowel duration: according to this interpretation, stressed vowels are longer, and, hence, more *compressible* than unstressed vowels. On this view, the hypothetical minimum duration would have to be imagined not as a hard boundary, but rather as a point at which segments are deleted rather than being subject to further gradual shortening. We will return to this idea in the course of our modeling work. Without access to the original data, no definite conclusion regarding diverging results on prominence and speaking rate is available; yet, it may be noted that the majority finding of stronger rate sensitivity of more prominent syllables also appears plausible given the observation that fewer syllables are *perceived* as distinctly prominent in fast than in slower speech (e.g. Crystal and House 1990).

Interestingly, Janse et al. (2003) refer to H&H theory as an explanation of their results,

arguing that stressed vowels shorten less strongly than unstressed vowels so as to preserve the informationally most important parts of the signal. An alternative account could be proposed based on the above considerations about incompressibility: stressed vowels shorten *more* strongly than unstressed ones because they are longer and, hence, there is "more room" for shortening without jeopardizing perceptual clarity. As a priori hypotheses, both accounts are probably equally valid. As we said in the introduction to this work, the theory needs to be implemented in order to determine which prediction it really makes.

Some studies have investigated the influence of speaking rate on final lengthening at major prosodic boundaries. Weismer and Ingrisano (1979) analyze data from different renditions of the phrase "Bob hit the big dog" produced by three speakers of American English at a "conversational" and a fast speaking rate. The authors report the phrase-final word "dog" to be shortened by generally less than 20% on average in the fast relative to the conversational rate condition, compared to 30–50% rate-induced shortening in the other words in the test sentence. As for deriving conclusions about the interaction between final lengthening and speaking rate, this study is obviously limited in that segmental content and relative prominence levels of the individual words are generally not controlled, but the words "bob" and "dog" may be roughly comparable, and in any case, the large and robust difference lends confidence to interpreting the result. It suggests that the duration of stressed monosyllables in American English is affected less strongly by speaking rate in utterance-final than in non-final position.

The above-introduced study by Beckman and Edwards (1990) also investigates lengthening of vowels at major intonational phrase boundaries under "fast", "normal", and "slow" speaking rate. Results indicate that the absolute durational increase from "fast" to "normal" tempo is roughly identical in phrase-final and non-final stressed vowels, which would point to a *proportionally* greater effect of final lengthening in the fast condition, in accordance with Weismer and Ingrisano (1979)'s results. One speaker even shows greater absolute phrase-final duration in the "fast" than in the "normal" condition. Results are inconclusive for the "slow" condition due to large between-speaker variation. In addition to this, Beckman and Edwards (1990) is the only controlled study we are aware of that examines final lengthening in *unstressed* vowels at different rates, by measuring /ə/ durations in *poppa, posing* versus *pop, opposing* sequences. Results indicate that with the exception of one speaker, final lengthening in unstressed vowels behaves differently under rate variation than in stressed vowels; final and non-final vowel durations converge in progressively faster speech, indicating a proportionally weaker final lengthening effect at fast rates in unstressed vowels. One caveat is that the non-final baseline in these materials is, in fact, phrase-initial, which means that it may be subject to prosodic timing effects as well.

Berkovits (1991) investigates final lengthening in slow and fast versions of read sentences in Hebrew, comparing target words in utterance-final and "phrase-final" position. In the latter case, the finality is with respect to some syntactic phrase. Berkovits reports greater target word duration in utterance-final than in phrase-final position in the fast condition, whereas in the slow condition there is no evidence for utterance-final lengthening. This result is open to multiple interpretations. It might be that there are really two processes, utterance-final and phrase-final lengthening, which are not distinguished in slow, but only in faster speech. Another possibility is that the difference between rate conditions is a consequence of prosodic restructuring in the slow condition, such that "phrase-final" and "utterance-final" in the slow condition really denote the same kind of phrase boundary. Since no phrase-medial baseline condition is provided, it is impossible to decide between these hypotheses. The study is also limited in that only two speakers are included. Moreover, as target words are polysyllabic and results are stated in terms of word duration, the exact domain of the final lengthening, as well as possible differences between stressed and unstressed syllables remain unclear.

Cummins (1999) investigates the three-way interaction between accentual lengthening, phrase-final lengthening and speaking rate in an experimental study with four speakers of American English, eliciting speech materials with continuous rate variation. The general conclusion is that final and accentual lengthening combine additively at all but extremely fast rates, in line with other investigations of accentual and final lengthening cited above. As for the influence of rate on evidence for final lengthening, results are not clear-cut; all speakers show some tendency to proportionally increase final lengthening under increasing rate, but individual results show that this trend may be reversed at extreme rates. Cummins (1999)'s results, too, indicate that large between-speaker variation characterizes the interaction between positional effects and speaking rate.

Smith (2002) investigates final lengthening at different rates in American English. The author reports vowel duration measurements in stressed (and, presumably, pitch-accented) monosyllables in phrase-medial and final position in carrier sentences read by 15 speakers at a slow and a fast speaking rate. Average proportional final lengthening is shown to increase from 43% in the slow to 63% in the fast condition, and this trend is also consistently observed for 13 of 15 speakers. Thus, Smith (2002)'s results support the earlier findings by Weismer and Ingrisano (1979), suggesting that in contrast to what most studies have reported for prominence-induced lengthening, final lengthening increases proportionally in faster speech in American English. Cross-linguistic generalizations are obviously not possible without data from other languages.

Of particular interest for our work are studies that have investigated temporal correlates of global variation on the H&H scale, either by explicit instructions to subjects, or by

creating external conditions that prompt a more listener-oriented speaking style. The latter strategy in particular has been applied in various studies of *Lombard speech*, i.e., speaking in noise. As mentioned above, one of the most robust findings from these studies is that "clear", hyperarticulated or Lombard speech is characterized by decreased speaking rate. Moreover, some studies have explicitly looked at influences of variation on the H&H scale on suprasegmental timing contrasts. By hypothesis, one may expect increased prosodic contrasts in more hyperarticulated speech, and this is indeed reported in Fant et al. (1991b)'s preliminary study of text reading in Swedish at different reading modes. The authors observe increased stressed-unstressed syllable duration ratios in a "distinct", hyperarticulated rendition of a text compared to the "normal" reading mode. However, as this study examines nine sentences comprised of uncontrolled materials read by a single speaker, it provides anecdotal evidence at best.

Cutler and Butterfield (1991) investigate stressed and unstressed syllable durations in speech produced by ten speakers of British English, eliciting variation on the H&H scale by first prompting subjects to produce experimental materials in a "natural" way, then asking them to repeat the utterances under the impression that a listener in another room had misperceived the initial renditions. Cutler and Butterfield (1991) analyze word-initial stressed and unstressed syllables matched for segmental material and find that the proportional increase in duration from the baseline to the clear condition is *weaker* in stressed than in unstressed contexts.[8] One potential caveat is that the phonological context of the target syllables is not always matched across stress conditions.

Patel and Schell (2008) report measurements on speech data gathered in an interactive game task under different external noise conditions from 16 speakers of American English. They present results grouped according to grammatical function, indicating strong increases in syllable duration in "agents", "objects" and "locations" from quiet to noisy conditions, compared to minimal increases in syllable duration in "functors" and "modifiers". No exact information on prosodic labels is given, it may be assumed that content words such as "agents", "objects" and "locations" tend to be prosodically prominent, whereas function words such as "functors" and "modifiers" are not. Segmental composition of the target items, moreover, is not controlled, but the variety of materials lends some confidence to interpreting the result as tentative support for the hypothesis that durational contrasts related to prosodic prominence increase in more hyperarticulated, or, more precisely, lombard speech.

Cho et al. (2011) investigate the effect of prosodic focus on vowel and syllable duration in Korean in "casual" and "hyperarticulated" speech. Analyzing test word produced in

[8]The authors actually refer to *strong* versus *weak* syllables. Inspection of experimental materials suggests that this is entirely commensurate to the stressed/unstressed distinction in their study.

uniform contexts by eight speakers under explicit instructions to speak "casually" and "clearly", they find that focus lengthens all parts of a word it applies to, and that this effect is proportionally stronger in hyperarticulated than in casual speech. Interestingly, vowels may even be shorter in the focus than in the non-focus condition in "casual" speech. The study does not allow for conclusions as to lexical stress, as Korean is said not to have lexical stress. The general result does support the assumption that prosodic contrasts are enhanced in hyperarticulated speech.

Arciuli et al. (2014) investigate durational patterns of utterance-medial trisyllabic words in lombard speech obtained from 27 speakers of Australian English. They do not report individual duration measurements, but rather quantify durational contrasts within target words using the *Pairwise Variability Index (PVI)*, that is, the average difference between pairs of adjacent syllable durations, which are normalized to the respective pairwise means. The authors find divergent results for words with different stress patterns: words with initial unstressed syllables show a slight increase in durational contrast in the lombard compared to the quiet condition, whereas the opposite pattern is observed for initially stresses words. This study, too, offers limited potential for generalization, as segmental materials are not matched across stress conditions. Moreover, raw duration measurements would have been more informative than the reported PVI scores.

As for the interaction between H&H scale variation and lengthening at prosodic boundaries, Garnier et al. (2006) report a preliminary study on controlled materials read by a single speaker of French in quiet and noisy conditions. Results suggest that the proportional increase in duration from the quiet to the noisy condition is considerably greater in utterance-final than in utterance medial syllables, whereas word-initial syllables undergo the strongest lengthening in utterance-initial position. Both observations are consistent with the hypothesis that prosodic contrasts are enhanced in hyperarticulated or lombard speech, but single-speaker studies can of course only provide tentative insights. A recent study of controlled materials produced by three speakers of Slovak under increasing noise conditions (Beňuš and Šimko 2015) indicates very small durational effects in the first place, and inconclusive evidence regarding a possible interaction.

3.5.3 Summary

Changes in overall speaking rate interact in interesting ways with the local effects of prominence and position. As for experimentally-induced changes in overall speaking rate, where the rate variation may be traced back to time constraints rather than communicative requirements, the majority of studies we reviewed here suggest that durational contrasts related to prominence effects diminish in faster speech, whereas final

lengthening at major prosodic constituent boundaries seems to increase proportionally in faster speech. As for speaking rate changes induced by global variation on the H&H scale, results are inconclusive, perhaps because different methodologies (application of a lombard signal versus explicit instructions to speak clearly) have been applied in different studies. A general caveat is that phenomena related to rate and global H&H scale variation are subject to strong inter-speaker variation. Especially in the case of H&H scale variation, moreover, further empirical study on a larger sample of languages is necessary to reach firm conclusions.

3.6 Overall Summary

In this chapter, we have reviewed evidence pertaining to suprasegmental speech timing, from the categories of prominence effects, positional effects, constituent length effects and effects of external speaking conditions. Based on this review, we can now formulate a "requirements specification" of suprasegmental timing effects that the model to be developed in the course of this work should account for. We hypothesize that constituent length effects can be eliminated as an independent category, because the evidence for these effects can be reinterpreted more convincingly as an epiphenomenon of prominence effects, and we will substantiate this by any empirical investigation in the course of this work. We close the chapter by presenting a concise list of the most important effects and interactions, grouped by the three remaining categories.

- Prominence Effects
  - Prosodic prominence is used in many languages to signal important units in the speech signal. At least two linguistically relevant levels of prominence can be distinguished: *lexical* stress, which refers to the enhanced prominence of a syllable within a word, and *accent*, which refers to the enhanced prominence of a word within a larger prosodic unit. Both effects are marked by increased duration of the unit they apply to in many languages.
  - When the lengthening effects of stress and accent interact, the result is not simply additive lengthening: there is a cross-linguistic tendency for accent to lengthen stressed vowels (and hence, by hypothesis, syllables) proportionally more than unstressed vowels. This may not be the case in word-final position.
  - The syllable count of accented words impacts their durational characteristics: absolute and proportional accentual lengthening of the word as a whole as well as of the component syllables/vowels diminishes in longer words. Absolute syllable/vowel durations are also shortened as a function of word length in

accented words, in an asymptotic fashion. This effect appears to be stronger in stressed than in unstressed syllables/vowels.

- Positional Effects
  - Syllables are lengthened at the end of large prosodic constituents, such as phrases or utterances.
  - This lengthening effect seems to be progressive, stretching from the last stressed syllable of a prosodic constituent to its end, given that there is no word boundary intervening. Lengthening is stronger the closer a syllable is to the boundary, with the possibility that unstressed syllables are "skipped" by final lengthening.
  - in contrast to different degrees of prominence, prominence and positional effects do not enhance each other when they occur in combination; prominence contrasts seem to be proportionally weaker in constituent-final than in non-final position.
- Effects of overall speaking rate
  - Stressed vowels/syllables shorten more strongly than unstressed vowels/syllables as overall speaking rate is increased, possibly due to incompressibility.
  - Available evidence from American English suggests that final lengthening at major prosodic boundaries increases proportionally under increasing speaking rates in stressed vowels, whereas the opposite may be true for unstressed vowels.
  - Hyperarticulated/clear/lombard speech is characterized by decreases in overall speaking rate. Empirical evidence as to interactions with prominence and positional effects is inconclusive.

Chapter 4

Explanatory Accounts of Suprasegmental Speech Timing

4.1 Introduction

In this chapter, we shall review existing explanatory accounts of suprasegmental speech timing. We have discussed some explanatory models of speech phenomena already in Chapter 2, but none of these was concerned with suprasegmental timing effects. Here we will focus mostly on implemented models, but also include some conceptual models in the discussion, which so far have not been explicitly formalized. We will assess how well these models account for the suprasegmental speech timing phenomena discussed in the previous chapter, and, crucially, how well they fulfill the criteria for truly *explanatory* accounts, i.e., independent motivation of the mechanisms employed.

4.2 Review

4.2.1 Oscillatory Models

Oscillatory models of speech timing conceptualize the hierarchy of prosodic constituents in speech as an ensemble of periodic *oscillators* at different prosodic levels such as the syllable, the foot, or the phrase The fundamental periods of the individual oscillators represent the durations of the respective prosodic constituents. Oscillatory models hold that, being nested within each other, oscillators at different prosodic levels impose mutual constraints on their fundamental periods, which are modeled by *coupling functions* that modify the oscillators' natural frequencies. This way, interacting oscillators can

generate complex timing patterns, even though oscillators in isolation produce simple periodic movement. In particular, since coupling between adjacent levels can be asymmetric, it is possible to model hypothesized dominant timing influences of individual prosodic levels, such that oscillators at some prosodic levels may be more or less susceptible to depart from their natural frequency than others.

We will discuss the seminal work on coupled oscillator models in speech timing, O'Dell and Nieminen (1999) at some length here. The starting point for this work is the long-standing controversy about the isochrony hypothesis. As we said earlier, the originators of this hypothesis (Abercrombie 1967, Pike 1945) claimed that there are some units in speech which speakers tend to produce at temporally regular, or *isochronous* intervals. For the "stress-timed" languages, notably English, it was hypothesized that speakers attempt at producing stressed syllables at regular intervals. A second group, the "syllable-timed" languages, was believed to exhibit a tendency to place all syllable onsets at temporally regular intervals, regardless of syllabic stress (ibid.). Subsequent research (e.g. Dauer 1983 and references therein) falsified the strong form of this hypothesis; in particular, the duration of the inter-stress interval (ISI) was found to be well approximated by a linear equation of the form

$$I = a + bn \tag{4.1}$$

where I is ISI duration, n is the number of component syllables and a and b are regression estimates for intercept and slope, respectively (Eriksson 1991). Based on a cross-linguistic analysis, Eriksson reports that, while b is close to 100 ms regardless of the language under study, the value of the intercept term a varies in an interesting way, clustering around 100 ms in the "syllable-timed" languages Spanish, Greek and Italian, and around 200 ms in the "stress-timed" languages English and Thai.

O'Dell and Nieminen (1999) hypothesize that this pattern results from an interaction between two oscillators at the syllabic and the ISI level. When viewed in isolation, both would oscillate at constant frequencies, which would be tantamount to generating isochronous syllables and ISI. In reality, however, isochrony at both levels cannot be maintained, because ISI may contain different numbers of syllables. Simply put, both oscillators have to settle for a compromise as a result: they need to entrain to a stable pattern, such that the frequency of the faster oscillator is an integer multiple of the frequency of the slower one, for any value of n, the number of syllables in the respective ISI. O'Dell and Nieminen hypothesize that the distinction between "stress-timed" and "syllable-timed" languages could be a matter of relative dominance in the interaction between the two oscillators: a "stress-timed" language would be one in which the ISI

oscillator[1] is more "dominant", i.e., more reluctant to depart from its natural frequency whereas in a "syllable-timed" language, the syllabic oscillator would be the dominant one.

We make no attempt to introduce the complete mathematical apparatus utilized by O'Dell and Nieminen (1999), and instead refer interested readers to the original paper, where the formal model is fully developed. Suffice it to say that the "compromise" between the two oscillators is modeled by altering their natural frequencies ω_1 (ISI) and ω_2 (syllables) by a *coupling function*, H, that is added with opposite signs to both ω_1 and ω_2, so as to produce the actual frequencies $\dot{\theta_1}$ and $\dot{\theta_2}$:

$$\begin{aligned} \dot{\theta_1} &= \omega_1 + H(\phi_n) \\ \dot{\theta_2} &= \omega_2 - rH(\phi_n) \end{aligned} \tag{4.2}$$

H is a function of both the *average phase difference* (ϕ) between both oscillators and the number of syllables n to be assembled in the ISI in question, i.e., the number of periods of the syllabic oscillator to be nested within one period of the ISI oscillator. The relative dominance of both oscillators is controlled by the *relative coupling strength* parameter r: for $r > 1$, the influence of H is stronger on the syllabic oscillator ω_2 than on the ISI oscillator ω_1 (r being 1 for ω_1, by implication). This makes the syllabic oscillator less reluctant to depart from its natural frequency and hence generates "stress timing". For $r < 1$, the reverse applies, generating "syllable timing". Crucially, the period of the ISI oscillator as a function of the number of component syllables n can be expressed as

$$T_1(n) = \frac{r}{r\omega_1 + \omega_2} + \frac{1}{r\omega_1 + \omega_2}n \tag{4.3}$$

It is easy to see that equation 4.3 has the same form as equation 4.1: the period of the ISI oscillator is a linear function of the number n of periods of the syllabic oscillator nested within it, with a positive intercept term for $r > 0$. In particular, the two fractions in equation 4.3, which correspond to a and b in equation 4.1, only differ in the enumerator, being r for the a and 1 for the b term. From this, it follows that the relative coupling strength parameter r is equal to the ratio between intercept and slope ($r/1 = r = a/b$) in a linear equation of this type, and thus can be estimated from empirical data. The values of r calculated from Eriksson (1991)'s results would be approximately 1 (100/100) for the "syllable-timed" and 2 (200/100) for the "stress-timed" languages. This shows that it is possible to account for the difference between the two putative language classes by assuming a dominant syllabic oscillator ($r < 1$) for "syllable-timed" and a dominant

[1] O'Dell and Nieminen use the terms "stress oscillator" or "stress group oscillator". These terms may be preferable to "ISI oscillator", because the focus of the original idea was on periodicity of stressed syllable onsets, rather than on isochrony of ISI as a linguistic unit. We shall nevertheless stick with our term, for the sake of consistency.

ISI oscillator ($r > 1$) for "stress-timed" languages. Thus, O'Dell and Nieminen's model of interacting oscillators at the syllable and ISI level generates the empirically observed timing pattern of ISI duration as a function of the number of component syllables.

O'Dell and Nieminen (1999) emphasize the abstract nature of their model. Their elaborations suggest that it is not intended as a real-time production model of speech – equation 4.3 makes a prediction for an isolated period of the ISI oscillator only, and it is probably not straightforwardly adaptable to real-life utterances, i.e. sequences of ISI with different n, where the oscillators would need to entrain to different frequency patterns in real time. The authors discuss some modifications to the model. In particular, they suggest using a "stress function" that slows down the syllabic oscillator at a particular phase of the ISI oscillator, so as to fit the greater duration of stressed compared to unstressed syllables. They state that this would only add a constant value, so that the general form of quation 4.3 would remain valid. Similarly, O'Dell and Nieminen describe how different "syllable types", presumably referring to something like syllables of different complexity, can be incorporated, by introducing separate a and b for each syllable type, which would have to be summed throughout the ISI oscillator period. Finally, the authors discuss the possibility of including further hierarchical levels, hypothesizing oscillatory mechanisms at other levels of the prosodic hierarchy as well. They demonstrate that this would result in a straightforward generalization, leading to "an expression of the slowest (θ_1) oscillator which is a linear function of all the numbers of different subunits contained in it" O'Dell and Nieminen (1999:1078).

The model by O'Dell and Nieminen (1999) provides a highly ingenuous and elegant account of the empirical data reported by Eriksson (1991). The crucial question is of course, how realistic are the assumptions it is based on – as we said earlier, computational modeling can demonstrate that a putative mechanism is *sufficient*, but not that it is *necessary* for explaining a particular phenomenon. One may note that a linear equation of the form $I = a + bn$ is a very simple thing to obtain, and there are probably many mechanisms that would account for it. A possible alternative, suggested by Eriksson (1991) himself and discussed by O'Dell and Nieminen, is that the empirical pattern is generated by a simple concatenative process, with the intercept a reflecting the greater duration of stressed compared to unstressed syllables.[2] Thus, the difference between "stress-timed" and "syllable-timed" languages might simply be a matter of higher stressed/unstressed syllable duration ratios in the former, without any need to posit interacting oscillatory mechanisms. This is of course not a necessary interpretation either; it assumes that syllable durations are independent of n, but as demonstrated by

[2]This can be demonstrated as follows: assume that syllable durations are independent of n. Stressed syllable duration S is identical to ISI duration I for $n = 1$ and thus can be calculated as $S = a + b$. Unstressed syllable duration U can be calculated as $I - S$ for $n = 2$, hence $U = (a + 2b) - (a + b) = b$. Thus, a is the difference between stressed and unstressed syllable duration: $a = S - b = S - U$.

Eriksson (1991) and, in fact, by O'Dell and Nieminen's model itself, this need not be the case. The point is that both hypotheses are equally valid *a priori*, and that independent evidence needs to be considered to decide between them.

An argument for the "stressed/unstressed ratio" hypothesis is that it converges with differences in language structure: "stress-timed" languages are characterized by features such as complex syllables attracting stress and reduction phenomena in unstressed syllables to a greater degree than "syllable-timed" languages (Dauer 1983). These differences alone would predict higher stressed/unstressed duration ratios, and thus greater r values in "stress-timed" than in "syllable-timed" languages. Results by Fant et al. (1991a) lend preliminary support to this view: these authors report virtually identical coefficients for "stress-timed" English and "syllable-timed" French in regression analyses of ISI duration on the number of *phones* in the ISI. This indicates that the difference reported by Eriksson (1991) may be primarily a function of language structure, possibly of greater differences in syllabic complexity between stressed and unstressed syllables in "stress-timed" compared to "syllable-timed" languages.[3] O'Dell and Nieminen (1999) cite Delattre (1966)'s finding of higher stressed/unstressed duration ratios in French than in English as counter-evidence, but as Delattre himself acknowledges, this comparison is confounded by positional effects: "stressed" syllables in French are always constituent-final, whereas only some stressed syllables are constituent-final in English.

O'Dell and Nieminen (2001) provide a more compelling argument against the view that cross-linguistic differences in ISI duration are purely a function of language structure: they describe a regression analysis of ISI duration on the number of component syllables for a corpus of Finnish, yielding an estimate of 104 ms for the intercept term a. O'Dell and Nieminen (2001) argue that this value cannot be explained by the difference between stressed and unstressed syllable duration, which they report to be only 13 ms on average in this database. This is certainly correct, but aggregate statistics over a whole corpus do not provide much information about the exact reasons for this discrepancy. It is conceivable that oscillatory mechanisms are responsible for the pattern of results, but it could also be due to interactions with other timing phenomena. As the authors themselves state (O'Dell and Nieminen 1999), durational processes inside the ISI need to be investigated in order to assess the predictions of their model.

As for this question, the key feature of O'Dell and Nieminen's model is that it naturally predicts a constituent length effect at the ISI level. O'Dell and Nieminen (1999) report

[3] Fant et al. (1991a) do report markedly more "stress-timed" coefficients for Swedish. This, however, may have to do with the particular system of word accents in this language, and in any case, it is not clear why English patterns with French and not with Swedish under a stress timing account.

that the fundamental period of the syllabic oscillator is given in their model as

$$T_2 = \frac{a}{n} + b \tag{4.4}$$

Thus, the period of the syllabic oscillator – and, hence, syllable duration – decreases as a function of n, particularly strongly for large a/b ratios, i.e., in "stress-timed" languages. There is actually only one extreme possibility of generating no constituent length effect at all, namely for a perfectly "syllable-timed" language, where syllables are isochronous and hence $a = 0$. This is the case at least in the basic model, where the frequency of the syllabic oscillator is assumed to be constant throughout the period of the ISI oscillator. O'Dell and Nieminen explicitly cite findings on alleged constituent length effects in support of their model, stating that "it has long been established that in numerous languages compression does occur (in all syllables) as the number of syllables increases" (1999:1076). As we have seen in Chapter 3, however, the existence of such effects is far from "established", particularly at the ISI level.

This state of matters challenges the coupled oscillator model. One might of course posit a superordinate oscillator at some constituent level other than the ISI which shows more evidence for "compression" effects such as the NRU or the accented word. However, this would mean giving up on the model's explanatory power regarding Eriksson (1991)'s results on ISI duration.[4] Moreover, while the coupled oscillator model was probably not originally intended as a real-time production model, some of its appeal stems from the intuition that a continuously oscillating device models a continuous train of time points, namely stressed syllable onsets in an utterance. This intuition would no longer be valid if a superordinate oscillator was posited at the level of the NRU or the accented word, because utterances can generally not be exhaustively parsed into these units.

In O'Dell and Nieminen (2008), the authors state that "(b)y letting syllable frequency vary during the stress group (with first syllable slower) while simultaneously letting the relative coupling strength of the syllable increase without limit, an extreme case can be achieved with all of the extra duration of the stress group concentrated in the first syllable" (2008:183). The accompanying figure confirms that in this case, the model would predict no compression at all as a function of n, corresponding to the simple case of concatenating a stressed and shorter unstressed syllables. Thus, these additional assumptions would allow the model to generate a pattern with no constituent length effects at all, but one might ask what would be the benefit of invoking a mechanism as complex as interacting oscillators for generating such a simple pattern. Moreover, with

[4] In fairness, since Eriksson (1991)'s regression results come from aggregated data from corpora, the model prediction would probably remain *statistically* valid if an oscillator at the NRU rather than the ISI level were assumed, as both units overlap to a large extent.

this parameter setting, the relative coupling strength parameter r would obviously lose its explanatory value, especially regarding cross-linguistic differences.

A further unsatisfactory aspect of the O'Dell and Nieminen (1999)'s model is its treatment of prosodic prominence. As we have seen, greater duration of stressed compared to unstressed syllables is achieved by invoking a "stress function" that slows down the syllabic oscillator at a particular phase of the ISI oscillator. The motivation O'Dell and Nieminen (1999) give for using this technique is a post-hoc one – the stress function is included *because* stressed syllables are typically longer than unstressed ones – rather than suggesting a principled explanation as to *why* this is the case. Thus, the oscillatory model misses out on a well-established intuition, namely that the lengthening of stressed syllables is connected to *perceptual* prominence, as detailed in Chapter 3. One might of course motivate the stress function by arguing that the syllabic oscillator is slowed down at a specific phase so as to meet the perceptual requirement of longer duration, but there is no explicit perceptually motivated mechanism in the model that would account for this lengthening. In any case, it is not clear whether the stress function adds any explanatory value to the model. O'Dell and Nieminen report no such results in any of their papers, which suggests that the stress function is essentially a data fitting device with no explanatory value on its own. In the model we are going to advance in this thesis, prominence is modeled by a principled mechanism that is informed by results from speech perception research, and we will show that several effects of prominence on speech timing emerge automatically from this feature of the model.

Saltzman et al. (2008) present a model designed to combine high-level prosodic oscillators and task dynamics at the gestural level, in order to develop a unified account of effects of prosodic structure on articulation. Their basic hypothesis is that *planning oscillators* at higher prosodic levels drive temporal coordination at the gestural level. As for modeling experiments, they concentrate on the suprasegmental domain and focus on the predicted constituent length effect. Saltzman et al. (2008) report simulations showing that their model produces shorter syllable durations in a tri- than in a bisyllabic ISI, but identify the problem that identical syllable durations are predicted for a given ISI length. They attempt to remedy this situation by introducing a "temporal modulation gesture", which they use to slow down the syllabic oscillator for an individual period. While not mathematically identical, this technique is equivalent to O'Dell and Nieminen (1999)'s "stress function", and similar criticism applies to it: it improves the fit to empirical data, but seems to add no explanatory value to the model, being motivated post-hoc, rather than on independent grounds.

The main source Saltzman et al. (2008) cite for the ISI triggering a constituent length effect is the corpus study by Kim and Cole (2005). This prompts them to introduce

a further modification to their model, as Kim and Cole (2005) report shortening as a function of the number of syllables in the ISI only for stressed, but not for unstressed vowels. Leaving aside the question whether constituent length effects in general are real this is debatable: authors of other corpus studies (Bouzon and Hirst 2004, Campbell 1988) do report shortening also in unstressed vowels, and Kim (2006), who provides a more complete discussion of Kim and Cole (2005)'s data, reports significant shortening of unstressed vowel duration as a function of the number of *phones* in the ISI. These observations strongly suggest that shortening in unstressed vowels may just have been masked by noise from other durational processes in Kim and Cole (2005)'s analysis. Saltzman et al. nevertheless introduce another modification to their model in order to accommodate Kim and Cole (2005)'s result: they modulate coupling strength ratio as a function of ISI oscillator phase, effectively switching coupling between the oscillators off during the unstressed part of the ISI, so that the unstressed syllable cycles are no longer affected by the number of syllabic cycles within the current period of the ISI oscillator. This technique has an even stronger ad-hoc flavor than the "temporal modulation gesture"; it quite obviously "hardcodes" the pattern observed by Kim and Cole (2005) into the model, without any motivation that is independent of the empirical finding itself. O'Dell and Nieminen (2008:183) are quite right in pointing out that what Saltzman et al. (2008) achieve is merely "to fit the model to empirical data".

A device similar to Saltzman et al. (2008)'s temporal modulation gesture has been applied to the modeling of boundary-adjacent lengthening in earlier work in Articulatory Phonology by Byrd and Saltzman (2003). They propose a "π-gesture" that is not directly related to any articulatory movement, but slows down the time flow during utterance production at prosodic boundaries. π-gestures are implemented using half-cosine functions, similar to the activation intervals of constriction gestures (i.e., those gestures that instantiate actual articulatory movements) in this approach. The π-gesture thus reaches its maximum amplitude at the boundary location and tapers off in a smooth fashion to both sides of the boundary. The magnitude of the temporal modulation is proportional to the gesture's amplitude at any point in time. The authors state that this device, beyond the obvious lengthening it triggers, reproduces specific characteristics of boundary-adjacent lengthening in speech: articulatory gestures to both sides of a boundary are lengthened; gestures overlap less in the vicinity of prosodic boundaries than elsewhere; boundary-adjacent lengthening may affect gestures which are not directly adjacent to the boundary, but the lengthening is progressive and affects gestures that are directly adjacent to the boundary more strongly than farther removed ones.

One may wonder here as well about the independent motivation of the modeling technique. What external evidence is there to support the assumption that boundary-adjacent lengthening is equivalent to slowing-down of the overall time flow in utterance

production? For example, the prediction that boundary-adjacent lengthening is progressive follows quite obviously from the decision to model the π-gesture with a half-cosine shape, and as far as we can judge, this decision is supported only by an argument from parsimony, namely Byrd and Saltzman (2003)'s reasoning that the π-gesture should have similar mathematical properties as articulatory gestures. Yet, as discussed in Chapter 2, the interpretation of boundary-adjacent lengthening is not straightforward, hence a more speculative approach may be acceptable here. In any case, one problem for the π-gesture approach is Turk and Shattuck-Hufnagel (2007)'s finding that phrase-final lengthening may affect multiple non-adjacent locations in polysyllabic words. As discussed by Turk and Shattuck-Hufnagel (2007), this would be problematic at least for the assumption of a single π-gesture at a prosodic break. The problem may be rectified by assuming multiple π-gestures at the locations affected by final lengthening, which, again, would be a rather ad-hoc construct.

Barbosa (2007) presents an oscillatory model of speech timing in Brazilian Portuguese. This model differs in some respects from the ones proposed by O'Dell and Nieminen (1999) and Saltzman et al. (2008). In particular, it seems to be geared more strongly towards descriptive purposes such as implementation in speech synthesis applications; Barbosa actually reports fitting the period of the syllabic oscillator to empirical data and discusses re-synthesis experiments with durations generated by the model. Yet, the model rests on quite strong assumptions about speech production, and Barbosa claims that his model provides a cognitively plausible account of speech timing.

Barbosa's model features a syllabic and a *phrase stress* oscillator. The basic organizational unit in this model thus seems to be some kind of prosodic phrase that is delimited by phrasal rather than lexical prominences, in contrast to the models reviewed above. Moreover, implementation details differ, but in any case, the same basic assumption applies: asymmetrical coupling between the phrase stress and the syllabic oscillator leads to a more or less pronounced constituent length effect at the phrasal level, and Barbosa explicitly refers to the distinction between stress timing and syllable timing as motivation for this feature of the model. We are not aware of dedicated studies of constituent length effects in phrasal units in Brazilian Portuguese, but given the empirical status of such effects in general, as discussed in Chapter 3, skepticism may be warranted concerning this prediction of the model.

In addition to the alleged constituent length effect at the phrasal level, Barbosa claims that the model also provides an account of final lengthening: it features a durational decay term that shortens syllabic cycles as a function of previous cycle durations, and this term not present between a phrasal stress location and the end of a phrase. Hence,

cycles of the syllabic oscillator occurring after the phrasal stress position will be lengthened relative to others. Barbosa states that the decay term is instantiated "in order to simulate the extended period of decay before duration increase towards phrase stress position" and claims that this mechanism "would explain final lengthening" of the syllables after the phrasal stress position (2007:733). These explanations strongly suggest that the decay term is an ad-hoc functionality, similar to Saltzman et al. (2008)'s "coupling strength modulation", whose purpose is to improve the fit to empirical data. This is of course completely acceptable for a descriptive model, but as there is independent motivation neither for the decay term itself, nor for the decision not to include post-accentual syllable cycles in it, we would argue that it does not provide a very convincing explanation of final lengthening. Even if such considerations are left aside, problems remain: the account of final lengthening as "absence of a pre-accentual decay term" would appear to predict that if the accent happened to occur early in the phrase (as for example due to an early contrastive focus), final lengthening would apply to all following syllables and thus effectively to most of the phrase. This is rather unrealistic.

Despite these reservations, there are a number of interesting aspects to Barbosa (2007)'s proposal. For example, in contrast to the models reviewed above, it makes use not only of frequency but also of amplitude parameters of the oscillators. While Barbosa reports no such attempts, utilizing oscillator amplitude could open up possibilities for more satisfactory modeling of prosodic prominence and for generating predictions with regard to different prominence levels. One interesting modeling result, finally, is reported on simulations showing that a linear increase in syllabic oscillator period leads to a sudden increase in variability and skewness of output duration distributions at some threshold value. Barbosa interprets this outcome with regard to experimental results on temporal synchronization in humans. He refers to experimental findings on the synchronization of finger tapping with a periodic external stimulus in humans, showing a similar dramatic increase in timing variability at a particular inter-stimulus interval, which happens to coincide closely with the value of the syllabic oscillator period at which the increase in variability occurs in Barbosa's model. Barbosa takes this finding to suggest that his model reproduces an temporal alleged boundary between "analytical" and "holistic" processing of time intervals in humans (2007:736). As a general comment, the author's attempt to devise a model that is both fit for usage in speech technology applications and motivated by cognitively plausible principles is certainly to be applauded. The available evidence just seems to suggest that interacting oscillators at different levels of the prosodic hierarchy *are* not a very plausible model of speech timing, at least as far as unconstrained speech production is concerned.

Rusaw (2013) presents an approach to the modeling of speech timing that combines oscillators and neural networks. In this paradigm, artificial neurons that emit single

oscillatory pulses of different periods are connected via excitatory and inhibitory connections, as is common in neural network modeling. In Rusaw (2013)'s approach, the individual neurons represent events at different levels of the prosodic hierarchy: phrasal boundaries, accents and syllables. Syllabic durations are represented by the time intervals between pulses emitted by the syllabic neuron. This neuron in isolation will emit pulses at regular periods; lengthening, for example due to prosodic prominence is modeled by an inhibitory connection from the accentual oscillator neuron, which delays the firing of the syllable neuron. Rusaw's motivation is to provide a link between the rather abstract concept of prosodic oscillators and their actual implementation in terms of neural structures in the human brain. She evaluates the model by comparing predicted durations to some English and French utterances. Based on visual comparison of graphs, she argues that the model delivers accurate predictions of syllable-level timing patterns. Yet, Rusaw reports that the parameters of the oscillator neurons were actually hand-fitted to match the empirical data, so it is not clear from these simulations what the neural oscillator paradigm really explains.

Rusaw also reports simulations of results on utterance-final lengthening by Turk and Shattuck-Hufnagel (2007), who found final lengthening to affect unstressed syllables directly adjacent to the boundary as well as stressed syllables earlier in the word, but not intervening unstressed syllables. Rusaw also replicates this finding in an empirical study of her own. In her simulation of these data, Rusaw incorporates an additional assumption: the phrasal neuron exerts excitatory influence on the accent neuron. This, in turn, causes the accent neuron to exert a stronger inhibitory influence on the syllabic neuron, so that syllabic pulses are delayed even more strongly in the presence of activation of both the phrasal and the accentual neuron. Using this configuration, Rusaw's model successfully reproduces the pattern observed by Turk and Shattuck-Hufnagel (2007): there is some lengthening is penultimate and antepenultimate stressed syllables, and strong lengthening in unstressed syllables directly adjacent to the boundary, where the cycle of the phrasal oscillator neuron reaches its peak. The really interesting result, in our opinion, is that in the case of an antepenultimate stressed syllable, the model predicts no final lengthening at all, although Rusaw's diagrams suggest that it is also under the influence of the phrasal oscillator peak.

As a general comment, Rusaw (2013)'s approach is quite interesting. However, it is difficult to assess the explanatory potential of the neural oscillator model, as little technical detail is provided on its implementation, and most of the evaluations reported in Rusaw (2013) focus on the model's ability to approximate raw syllable durations, rather than on its potential to provide principled explanations of durational patterns. Rusaw seems to imply that not only relative syllable durations, but also locations of accents and phrase

boundaries are emergent properties of her modeling paradigm, but due to her above-mentioned technique of hand-fitting oscillator parameters, it is not clear to what extent this statement is really justified. In any case, Rusaw's modeling experiments on final lengthening are encouraging, and it would be interesting to see more in-depth studies of prosodic interactions in the model. On a final note, it appears that the neural oscillator model does not necessarily predict compensatory timing relationships between prosodic levels, as seems to be the case with other oscillatory approaches, at least as long as no additional assumptions are imported.

The critical points raised earlier about coupled oscillators that do predict such relationships as a general account of prosodic timing in speech are not meant to suggest that such models are not useful for modeling speech production under special circumstances. This, for example, seems to be true of speech produced in so-called *speech cycling* experiments. In the speech cycling paradigm, subjects have to entrain repeated productions of speech utterances with alternating periodic sequences of high and low metronome tones, resented in different phasing relations. Evidence from this paradigm (Anbari et al. 2013, Cummins and Port 1998) demonstrates the emergence of "rhythmical attractors", i.e., speakers tend to place the prominences in their productions at certain phases that divide the repetition cycle defined by the low tones into harmonically spaced intervals, even if the phasing of the intervening high tones is different. Saltzman et al. (2008) show that it is possible to account for this result using an ensemble of oscillators, where two pairs of nested oscillators with bi-directional coupling represent the external stimulus and the production of the subject, respectively, and the superordinate and the subordinate oscillator from the former pair also exert unidirectional coupling on their counterparts from the latter pair of oscillators. Moreover, Tilsen (2009) in an investigation of EMA data from a speech cycling experiment observes a correlation between temporal variability at the phrasal and gestural level, and shows that a collection of hierarchically organized oscillators correctly predicts this outcome.

Cummins and Port (1998) state that effects observed in the speech cycling paradigm are task-specific and may not generalize to unconstrained speech production. However, there may be more "natural" situations where speech production is subject to comparable entrainment constraints, such as various forms of joint speech (Cummins 2009). Moreover, oscillatory approaches may provide a suitable paradigm for modeling timing relations between the utterances of interlocutors in conversation. For example, Włodarczak (2014) in a study of the timing of overlaps in conversation observes that overlapping turns tend to be initiated at salient points in the interlocutor's speech, such as syllable boundaries, and suggests that this pattern may be captured by modeling with coupled oscillators. Indeed, oscillatory models have been applied with some success to tasks such as the prediction of the timing of feedback utterances in dialogue (Wagner

et al. 2013). These results suggests that periodic oscillators may be a useful ingredient of models of the temporal entrainment between interlocutors in conversational situations. Given the available evidence, it just appears that as a model of (within-speaker) prosodic structure in unconstrained speech production, a full hierarchy of interacting oscillators may be rather too strong an assumption. In one of the subsequent chapters of this work, we will try to substantiate this preliminary evaluation of oscillatory models by a detailed empirical investigation of their main predictions.

4.2.2 The Converter/Distributor Model

The Converter/Distributor (C/D) model, first introduced in Fujimura (1994), is a fully-fledged computationally implemented theory of the phonetics-phonology interface. As such, it covers considerably more than suprasegmental speech timing, but we will focus our discussion on those aspects of the model that relate to suprasegmental speech timing. In any case, prosody occupies a quite central position in this approach: the C/D model maintains that syllables are the basic organizational unit in speech. The syllabic structure of an utterance is represented in this model by a pulse train. The amplitude of individual pulses roughly corresponds to prosodic prominence. In this model, syllable "base durations" are derived by constructing *syllable triangles* with the syllable pulses as center lines, as shown in Figure 4.1. This is done by specifying a fixed value for the apex angle, referred to as *shadow angle* in the C/D model (Fujimura 2011). The *converter* module of the model derives these apex angles from metrical representations of utterances. The syllable "base" duration (not necessarily equal to, but correlated with acoustically measured syllable duration) is equal to the triangle base width. In addition to the syllabic pulse train, the model features additional tiers, e.g. for consonantal and vocalic gestures. The points at which the legs of a syllable triangle touch the base instantiate pulses on the consonantal tier, representing onset and coda consonants. The actual consonantal gestures are modeled by low-pass filter impulse response functions to these pulses. Simultaneously, the syllable pulse excites a tongue body gesture for the vocalic center of the syllable, likewise modeled by a damped impulse response function. The derivation of actual gestures from an abstract syllable pulse sequence is referred to as the *distributor* module of the model.

The core interest of the C/D model lies on prosodic effects on articulation. We will not discuss predictions of the model concerning these effects, but rather focus on what it has to say about suprasegmental timing in the narrow sense. To begin with, the shadow angle technique provides an elegant account of the relationship between prosodic prominence and syllable duration: since the shadow angle is assumed to be fixed, a syllable

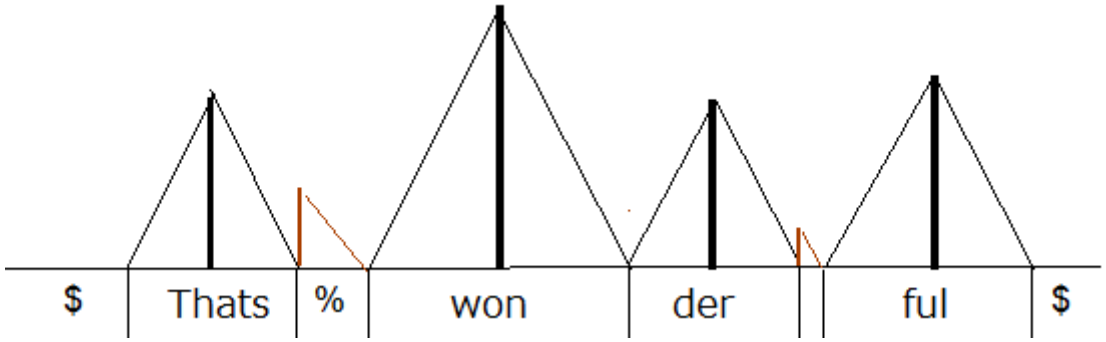


FIGURE 4.1: Syllable triangles representing an utterance in the C/D model (from Fujimura and Erickson (2004). The "%" and "$" symbols represent speech pauses at prosodic boundaries of different strengths.

with a greater pulse magnitude will also have a greater base duration.[5] In particular, the trigonometric derivation of the base duration guarantees a linear relationship between duration and prosodic prominence. Studies of the acoustic correlates of perceived prominence (Fant and Kruckenberg 1989, Portele 1998, Portele et al. 2000) have generally found linear relationships as well. This is in line with the C/D model prediction, although one may argue that the syllabic pulse magnitude in the C/D model actually reflects something like *intended*, rather than *perceived* prominence, and it is not clear whether both concepts can be equated with each other.

Perhaps more interestingly, one may speculate about the treatment of speaking rate effects on suprasegmental timing in the C/D model: changes in overall speaking rate are incorporated in the C/D model by changing the shadow angle for a given utterance, so that smaller or greater base durations are derived for a given magnitude. This technique would appear to predict that duration ratios between stressed and unstressed syllables remain unchanged under rate variation, which is at odds with most empirical findings reviewed in Chapter 3. Reproducing the predominantly observed pattern of stronger rate sensitivity of more prominent syllables would presumably require to change the shadow angle on a syllable-by-syllable basis, depending on the pulse magnitude of each individual syllable rather than uniformly for a whole utterance. Alternatively, one might conceive altering syllable magnitudes themselves in order to achieve rate variation, which would also have to be sensitive to the absolute magnitude of a given syllable in order to simulate the interaction between prominence and rate. Either strategy, however, would seem relatively ad-hoc.

The C/D model features an extra category of "half-triangles" for representing speech pauses. As for lengthening of syllables before pauses, we are not aware of explicit

[5] Fujimura (2011) states that in later versions of the model, different shadow angles are assumed for different "syllable types", e.g. light vs. heavy syllables.

proposals for implementation within the C/D paradigm. An obvious choice, again, may be to manipulate the shadow angle for boundary-adjacent syllables in order to instantiate boundary-adjacent lengthening. This would enable the model to simulate lengthening of syllables with different degrees of prominence, in contrast to a conceivable alternative technique of modeling boundary effects by manipulating syllable magnitude, which, in effect, would render all boundary-adjacent syllables prominent. Incorporating final lengthening by manipulating the shadow angle would predict final lengthening of similar proportional magnitude in stressed and unstressed syllables, which is in accordance with some of the results reviewed in Chapter 3.

(Fujimura 2011) explicitly states that the C/D model does not feature any components which would evoke tendencies towards periodicity. Given the available evidence reviewed in Chapter 3, this appears to be a reasonable modeling decision. As we saw, one well-supported effect that appears to minimally require interaction between the syllabic and higher prosodic levels is the inverse relationship between syllable duration and syllable count in accented words that has been observed in many languages. Whatever the explanation for this effect, it is not clear how it could be borne out by the C/D model without explicitly manipulating shadow angles or, perhaps, syllable magnitudes according to the syllable count in a subset of an utterance that is defined as an accented word. Maybe this effect – and others, such as the reported greater accentual lengthening of stressed compared to unstressed syllables – could arise automatically if an additional prosodic level featuring extra "accent pulses" and, likewise, "word triangles" were incorporated. As extensions of these kinds have not been implemented, one can of course only speculate whether they would reproduce the observed effects in a non-trivial fashion.

As a general comment, we would like to express our appreciation of the C/D model as a computationally explicit theory of phonetic implementation. The points we have raised in the foregoing decision may not be entirely fair, because the focus of the C/D model does not lie on the purely suprasegmental timing effects we are interested in. Our highlighting of phenomena for which the C/D model fails to make correct predictions is therefore not meant to be dismissive of this model, but, ultimately, to underline the fact that an explanatory account of these phenomena is still a desideratum.

4.2.3 Other Approaches

In the remainder of this chapter, we will discuss some explanatory accounts (not necessarily implemented models) we regard as interesting but less central for our argument, because they focus on isolated phenomena and/or have not been implemented. These accounts are mostly concerned with constituent length effects in speech. The exception

to this is Lindblom (1968), who sketches an explanatory account of final lengthening. His proposal builds on the intonation model by Öhman (1967), which incorporates an explicit account of subglottal pressure in speech production by means of a "physiological intensity contour", which "is constant during the beginning of an utterance but then falls towards the end" (Lindblom 1968:4). Lindblom additionally assumes that the amount of energy expended per syllable is roughly constant except for variations induced by prominence, and can be approximated by integrating the physiological energy contour over a syllable's duration. Final lengthening would then follow from declination of the physiological intensity contour towards the end of an utterance – syllables would have to be lengthened in order to maintain approximately constant energy.

This proposal is certainly elegant and provides a welcome independent explanation for final lengthening phenomena in speech. The assumptions of (1) constant energy expenditure per syllable and (2) declining "physiological intensity" towards the end of an utterance may be debated – Oller (1972) argues against the latter assertion, based on the observation that speech amplitude does not usually drop towards the end of utterances, although it is not clear what data his statement is based on. Declination in subglottal pressure over an utterance has been observed for example by Strik and Boves (1995) and Trouvain et al. (1998), which would support Lindblom's approach. On the other hand, a final lengthening mechanism based on subglottal pressure declination would probably predict gradual deceleration over the course of an utterance. This is at variance with empirical findings on final lengthening, which is indeed progressive, but tightly localized to linguistic entities, and, furthermore, may affect discontinuous locations in an utterance (Turk and Shattuck-Hufnagel 2007).

Lindblom et al. (1981) propose an explanation of constituent length effects at the word level in speech based on hypothetical constraints on short-term memory. They hypothesize that speech planning involves a short-term memory buffer for word-level elements, into which the phonetic plans for syllables have to be loaded. Since the buffer is of limited size (but expandable), phonetic plans have to be executed in shorter time as more of them are loaded into the buffer; compressibility constraints, inspired by the descriptive model by Klatt (1973) ensure that duration compensation is only partial. This way, the asymptotic nonlinearity of the shortening effect found in the empirical data is borne out. The model also provides an account of final lengthening phenomena: "units that have been processed (preceding syllables) call for less space saving than units that are on their way to be processed (following syllables)" (Lindblom et al. 1981:72).

Lindblom et al. (1981) do not present an actual implementation of their model – rather, they fit descriptive formulas to empirical data and assign interpretations to these. The model is obviously strongly influenced by metaphors from computing, as is evident

from the usage of concepts such as storage space, memory, or executions. Lindblom et al. (1981)'s proposal is quite ingenuous, but assuming that speech timing results from things such as elastic memory buffers or recursive shortening operations at the planning stage is of course a rather strong conjecture. If one accepts White (2002)'s hypothesis that constituent length effects at the word level are simply an epiphenomenon of word prominence, moreover, Lindblom et al. (1981)'s short-term memory account would be rendered rather obsolete.

An interesting proposal for explaining the observed inverse relationship between word length and syllable duration has been put forward by Nooteboom (1985). Nooteboom rightly observes that "isochrony" principles explain nothing by themselves; once they are invoked, the question remains why speakers should tend to keep the durations of certain units constant. Nooteboom' proposal draws on the proposed relationship between speech timing and information redundancy that we reviewed in Chapter 2. He assumes that "(p)recise articulation costs time", and that "(w)hen segmental information can be confidently predicted from context, the speaker can allow himself to speak sloppy and fast" (1985:244). These formulations are clearly reminiscent of efficiency-based accounts. As for the influence of word length on segmental duration, Nooteboom maintains that segments in longer words are more redundant and, hence, can be shortened: "in general, we need all phonemes in their correct order to recognize monosyllabic words like CAN. A polysyllabic word like ELEPHANT, however, is, for instance, uniquely determined by the initial word fragment ELEPH..." (1985:245). This, according to Nooteboom, would predict shortening, especially of the redundant part of the word. As Nooteboom himself notes, this prediction is not borne out: segments are commonly *lengthened* towards word endings. He invokes an additional principle to accommodate this observation: speakers strive to "provide the listener as quickly as possible with the auditory cues necessary for initial recognition" (Nooteboom 1985:246), whereas they can relax the efficient information transmission requirement after a word has been recognized, which often happens well in advance of its end.

Although Nooteboom (1985) acknowledges that his proposal may miss some generalizations, we still think that it is well-motivated. The model we are going to propose will include some features compatible with Nooteboom's ideas, although there will be no direct consideration of phenomena such as the time course of word recognition, which is important for Nooteboom's approach. In any case, it would be interesting to see an implementation of Nooteboom's ideas, and we will supply some speculations as to a model architecture that could accommodate similar mechanisms in the final chapter of this work.

Fujimura (1987) describes a model that represents speech utterances as arrangements of linearly concatenated springs between two hard boundaries, with the individual springs referring to segmental units. Fujimura also discusses an extension of this model that features springs at different prosodic levels, which are nested within each other. While he does not explicitly uses this term, this model is very similar to the oscillatory models discussed above, and the mathematic bases of both paradigms are probably compatible. Timing effects could be modeled in this framework by manipulating either the stiffness of individual springs, or the external forces acting upon them. Fujimura (1987) discusses evidence from temporal alignment between certain articulatory landmarks in spoken utterances with identical segmental content but different accent placement. The analysis reveals that there are "time domain(s) over which a uniform difference in utterance speed is observed' (1987:118), which he takes to support his modeling assumptions. Since Fujimura's model has not been implemented, it is, however, not clear whether this prediction is actually borne out. While his proposal is certainly interesting, evaluation has to remain speculative in the absence of a working implementation.

Messum (2008) proposes a unified explanation for some speech timing phenomena based on developmental properties of speech breathing in humans. He maintains that, while adults' speech is characterized by the capability of "inflating the lungs and then largely speaking on relaxation pressure" (Messum 2008:2409), so that larger stretches of speech can be produced in one breath group, this is different in children: due to the greater compliance of their lung and chest wall tissue, children have to adopt a more pulsatile style of speech breathing, with each syllable corresponding to a respiratory pulse. Messum (2008) claims that this feature of speech breathing in children is the source of two segmental timing effects, shortening of vowels before voiceless as opposed to voiced consonants, and before consonant clusters as opposed to singleton consonants: the temporal extension of a breath pulse has to be shared out between the vowel and a longer consonant (in the case of the voiceless/voiced opposition), or, respectively, between the vowel and more consonants (for the singleton/cluster oposition). Messum (2008) states that these patterns remain as speech breathing becomes more adult-like. Importantly, he claims that a similar explanation pertains to "foot-level shortening", i.e., a constituent length effect at the ISI level. Messum (2008) maintains that children in stress-accent languages produce prominences by increased respiratory system drive. He further states that the typically reduced unstressed syllables in these languages offer very little aerodynamic resistance; hence, the ISI in these languages is produced in children's speech as a "single high-resistance unit" (2008:2412), so that, by extension, the same compensatory timing relation observed at the syllabic level in the case of consonant clusters applies also at the ISI level. Again, Messum (2008) claims that this effect is fossilized in the speech of adults.

Messum (2008)'s proposal is quite ingenuous, and we applaud his attempt to derive physiological explanations for speech timing phenomena. One problem with his account of the segmental timing phenomena is that at least the effect of postvocalic voicing on vowel duration operates across syllable boundaries, as is shown by evidence from German (Braunschweiler 1997). Yet this is a minor problem, and it could be circumvented by positing the *vowel-to-vowel interval* as the domain of breath pulses. As for the explanation of "foot-level shortening", however, Messum (2008)'s proposal obviously requires quite strong additional assumptions, namely the identification of the ISI as a "single high-resistance unit", which, as far as we can see, rests on conjecture. It appears to predict that shortening is observed mostly in the unstressed part of the ISI, which is at variance with some empirical findings (Kim and Cole 2005). Finally, the possibility has been raised that foot-level shortening is, in fact, entirely spurious and would thus not require any explanation at all. We will supply empirical evidence for this proposition in a later chapter of this work.

Finally, there has been one recent theoretical proposal for a comprehensive optimization-based model of speech timing including suprasegmental aspects. This proposal has been put forward by Turk and Shattuck-Hufnagel (2014b) in the context of the debate on *extrinsic* versus *intrinsic* timing in speech. This debate has been sparked by Fowler (1980)'s review of previous models of coarticulation. Fowler (1980) states that the failure of these models to account for observed facts about coarticulation is due to the fact that they assume *extrinsic* timing, i.e., they presuppose that speech is planned as a sequence of atemporal discrete segments, and that temporal structure is just superimposed on this structure during execution. Fowler's alternative proposal of intrinsic timing in speech production has paved the way for the gestural account of Articulatory Phonology (Browman and Goldstein 1992), which views physically instantiated articulatory gestures – and, hence, objects with an inherent temporal dimension – as the basic building blocks of speech. Turk and Shattuck-Hufnagel (2014b) set out to challenge this view and propose a model of extrinsic speech timing, in which a central "timekeeper" plans the phonetic execution of symbolic linguistic structure.

Turk and Shattuck-Hufnagel (2014b) envision the timekeeper in their model to be guided by optimization principles, trading off timing and accuracy requirements in articulation against movement costs. These costs are balanced against each other by the hypothesized goal of speakers to achieve an even distribution of recognition likelihood across an utterance, which we already encountered in the discussion of Aylett and Turk (2004) and related works in Chapter 2. Thus, it may be assumed that movements costs would be prioritized in highly predictable environments, whereas faithful realization of phonetic targets would be prioritized in less predictable environments. Turk and Shattuck-Hufnagel (2014b) stipulate that coarticulation effects should be an automatic consequence of these

trading relations in their model. Since Turk and Shattuck-Hufnagel's proposal has not been implemented, this statement remains somewhat speculative.

4.3 Discussion

Our review of existing explanatory models of suprasegmental speech timing leads us to conclude that an explicitly formalized comprehensive explanatory account of suprasegmental speech timing is still lacking. The coupled-oscillators paradigm and Fujimura's C/D model come closest to meeting this requirement. The coupled-oscillators paradigm, in our opinion, is founded on rather questionable assumptions, at least as far as unconstrained speech production is concerned. The C/D model, although grounded in a plausible and explicit account of prosodic structure, does not focus on the class of suprasegmental timing effects and interactions that we reviewed in Chapter 3 of this work. It appears that many of the prosodic interactions discussed in Chapter 3, such as those between different levels of prominence, or between prominence and positional effects, have not been addressed in any explanatory model. The model we present in the subsequent chapter of this work and test in the following chapters represents an attempt at filling this research gap, by focusing on exactly these interactions.

Several of the accounts reviewed in this chapter focus on constituent length effects in speech. Our review of empirical results in Chapter 3 of this work suggests that such effects may in fact not require any dedicated timing mechanism, but are explicable as a mere corollary of word prominence, or, in other cases, statistical artifacts. The model to be developed in this work will therefore not feature any component dedicated to introducing constituent length effects (although such effects will fall out automatically from word prominence, as has been argued by White (2002)). After introducing the architecture of our model in Chapter 5, we will dedicate Chapter 6 of this work to an empirical investigation of the predictions made by the models reviewed in this chapter.

Part II

Model Definition and Results

Chapter 5

Model Definition

5.1 Introduction

We will now introduce the architecture of our own optimization-based model of suprasegmental speech timing, informed by the review in the preceding chapters. In Chapter 2 we have seen that efficiency-based approaches that trade off minimization of effort against maximization of perceptual clarity provide a well-suited explanatory platform for many phenomena in speech. In Chapter 3 we have provided a concise overview of speech timing phenomena in the suprasegmental domain that an explanatory model should account for. In Chapter 4, we have reviewed existing explanatory models of suprasegmental speech timing and argued that they fall short of providing a satisfactory account of the empirically observed phenomena in this domain. In the review of optimization-based explanations of speech patterns, the approach adopted in Šimko (2009)'s Embodied Task Dynamic (ETD) model in particular has emerged as a promising candidate for explaining timing phenomena in speech. In this thesis, we will adapt the general principles outlined in Šimko (2009)'s articulatory model to the modeling of timing phenomena at the suprasegmental level. In what follows, we will introduce the individual components of the model and supply external evidence to justify our modeling decisions. We will also briefly introduce the optimization procedure as such.

One fundamental simplification will be made in adapting the optimization model from Šimko (2009) to the modeling of suprasegmental speech timing: we will assume that suprasegmental timing phenomena in speech can be satisfactorily modeled in terms of acoustic durations of prosodic constituents. This allows us to exchange the differential calculus necessary to derive solutions for the mass-spring equations used in Šimko (2009)'s articulatory model for simple arithmetic on real numbers. This is obviously a massive abstraction from actual, physically instantiated speech, and we are certainly

not intending to suggest that speakers plan utterances by computing real numbers that stand for prosodic constituent durations.[1] Yet we conjecture that direct modeling of acoustic durations represents a reasonable first-pass strategy, combined with the advantage that it simplifies the model by orders of magnitude. As we shall see later, this modeling approach is sufficient for generating theoretically interesting predictions.

Optimization in the present model will thus be computed over vectors of real numbers that represent syllable durations in hypothetical speech utterances. These syllable sequences do not necessarily relate to any real-life utterances (although they could, in principle), but are generally rather abstract in nature. The reason for this is that we are not interested in a close approximation of the durational characteristics of any real-world dataset, but in deriving principled explanations for a narrow range of basic timing effects. Given this, we deem it the most appropriate approach to study these effects in isolation, and to ignore other sources of durational variation, such as syllable structure, adopting a *ceteris paribus* assumption (such as "all else being equal, stressed syllables are longer than unstressed syllables") in any statement about the processes we are investigating. The choice of syllables as the basic unit in the model is motivated by purely pragmatic considerations: syllables are an intuitive organizational unit in speech, and they represent the smallest independent unit in the suprasegmental domain, according to our definition introduced in Chapter 3. Moreover, the timing processes we are interested in are commonly envisioned to apply at the syllabic level, even though empirical investigations of these processes are frequently conducted on vowel durations. In any case, the fact that the model operates on syllables is not meant to raise any strong claims about these (rather than other) entities functioning as primary "processing units" or "units of speech perception".

To recapitulate, the general form of the model introduced by Šimko (2009) is as follows:

$$C = \alpha_E E + \alpha_P P + \alpha_D D, \tag{5.1}$$

where C is the overall cost to be minimized, E and P represent the hypothesized drives towards minimizing effort and maximizing perceptual clarity, respectively, D is related to overall speaking rate and α_E,α_P and α_D are scalar weighting factors that control the relative influence of the individual component cost functions. In our model, we will utilize the general form of Equation 5.1, although the individual components will naturally be re-defined. The three components of C are functions of syllable durations $S = [s_1, \ldots, s_n]$ of individual syllable durations s_i in a series of n syllables that represents

[1] A more thorough discussion of how our model relates to "online" processes such as speech planning will be provided below. To preempt this discussion, the short answer is that the model has essentially nothing to say about such processes.

a hypothetical utterance. A numerical optimization algorithm will be employed to find the vector S of syllable durations such that overall cost C is minimized.

One important feature to be implemented in the model is the assumption that trade-offs between minimizing effort and maximizing communicative success may apply both at a global and at a local level. We have seen this spelled out in the account of prosodic prominence as localized hyperarticulation, and also in the local modification of component cost functions in Šimko et al. (2014b) and Beňuš and Šimko (2014). In our model, we will systematize this assumption by adding local weighting factors to all three component cost functions, which modify their relative influence for individual syllables. These local weights will be introduced in the discussions of the individual component cost functions in the remainder of this chapter.

5.2 Model Components

5.2.1 Effort Cost E

Component cost E represents the hypothesized tendency towards minimizing effort in our model. It therefore requires an estimate of the physical effort necessary to produce a syllable as a function of its duration, as our model is restricted to the temporal dimension of speech. Thus, it is clear that component E in particular will require quite strong simplifications, as the model abstracts away from physically instantiated speech production. Nevertheless, we will be able to make some principled considerations, leading to an effort estimate that serves its purpose.

Following Howard and Messum (2011), we make the fundamental assumption that effort in speech comes from two sources, articulation and phonation: on the one hand, metabolic energy is needed to move the articulators towards their targets; on the other hand, the organism has to provide respiratory energy, so as to maintain the necessary airflow. Muscular activity is also necessary to sustain vibration of the vocal cords during voiced portions of the speech signal.

As for articulatory effort, we have seen that different measures, such as articulatory velocity, displacement of articulators, or the force acting on them, have been proposed. It is not possible to link any of these straightforwardly to durations of acoustic events, such as syllables, because speakers can choose between different articulatory strategies in order to vary duration. As we saw in the discussion of Lindblom (1963), shortening in particular may be instantiated by increasing the velocity of articulatory movements, or by undershoot of targets. If we once more imagine an articulator as a mass-spring

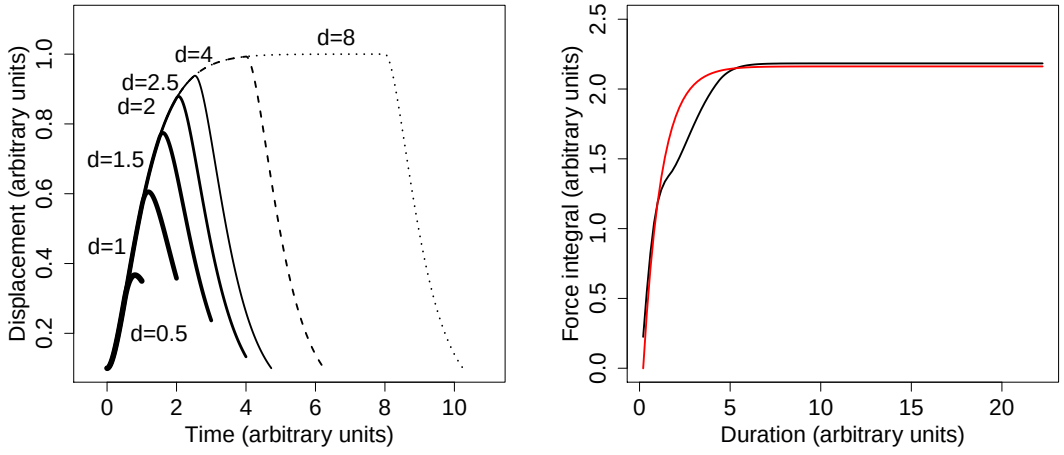
system as in Equation 5.2, the former would be instantiated by increasing the system stiffness K, whereas the latter would be an automatic consequence of a shorter activation interval, as shown in Figure 2.1 in Chapter 2. For this reason, we will make one simplifying assumption: we will assume that the undershoot case, i.e., no increased stiffness at shorter durations, is the default behavior of speakers. This is consistent with economy principles, i.e., the idea that high movement velocities are avoided (Howard and Messum 2011). We would argue that this is a reasonable default assumption, given that many studies have reproduced Lindblom (1963) finding of a correlation between spectral reduction (as a measure of target undershoot) and durational shortening (e.g. Aylett and Turk 2004, 2006, Hirata and Tsukada 2009, Moon and Lindblom 1994, Nadeu 2014, Van Son and Van Santen 2005).

$$M\ddot{z} = -K(z - z_0) - B\dot{z} \tag{5.2}$$

It turns out that with this simplification in place, we *can* come up with a principled estimate of articulatory effort as a function of syllable duration. This estimate was derived from a simple computational mass-spring model that, similar to the model employed by Kirchner (1998), simulates the movement of a single critically damped linear spring, interpretable as a single, "generic" articulator, towards a vertical target.[2] In this work, we follow Kirchner (1998) and Šimko (2009) in employing the *force integral*, i.e., the sum of forces acting on the model articulator over time, as an effort estimate. Higher effort is thus linked to greater forces, which in our opinion, is intuitively compelling. Force is defined as the product of mass and acceleration (the left-hand side of Equation 5.2), but for the case of a single spring, mass is just a multiplicative constant and can therefore be disregarded. Our estimate of articulatory effort as computed by the mass-spring model is therefore simply proportional to the absolute value of the acceleration of the spring integrated over time. The mass-spring model was repeatedly run, holding the values of its free parameters – vertical starting value and target, and, in particular, stiffness – constant, whereas the duration of the activation interval was incremented by a constant amount, so as to obtain an articulatory effort estimate as a function of duration.

The left panel of Figure 5.1 shows trajectories of the spring over time for different activation interval durations. The right panel of Figure 5.1 plots the value of our articulatory effort estimate, the force integral, in black as a function of the duration of the activation interval, and, hence, the gesture. The red line shows that a negative exponential function provides a reasonable approximation to the data ($R^2 = 0.91$), and as the left panel of Figure 5.1 illustrates, this is also intuitively plausible: increasing a short duration

[2]The implementation of the mass-spring model was kindly provided by Juraj Šimko. The source code (programmed in MATLAB) is included in Appendix B of this work.



(a) (b)

FIGURE 5.1: Panel (a): plots of displacement over time as computed by the mass-spring model, for some values of activation interval duration d. Parameters: stiffness=3, vertical target=1, vertical starting point=0.1. Fainter trajectories indicate longer activation intervals. Panel (b): force integral (articulatory effort estimate) as a function of activation interval duration as computed by the mass-spring model (black) and fitted negative exponential function (red).

is linked to increased displacement of the relevant articulator, as articulatory targets have not yet been reached. Lengthening a short syllable is therefore achieved by extra movement, and, hence, results in increased effort. At some duration, however, articulators reach their targets – in the left panel of Figure 5.1, this can be seen for activation interval duration $d = 4$. Lengthening the syllable beyond this duration is achieved by stretching the steady states of segments within the syllable (mostly the vowel), which does not require any articulatory movements, and is therefore free on the articulatory effort dimension. This intuition is well captured by the initial steep rise and subsequent plateau of our articulatory effort estimate.

It is obvious that this derivation of articulatory effort is a strong simplification. Besides assuming constant stiffness, it also ignores the fact that syllables usually contain several articulatory gestures, which may mutually influence each other, as is expressed by the anatomical linking in ETD. Yet, as stated above, in all model simulations to be reported in this work, we will assume a ceteris paribus condition, i.e., it will be assumed that syllables involved in any comparison will differ only on the factor that is being varied. All comparisons therefore have to be imagined as involving syllables that include the same gestures and differ only on the suprasegmental variables of interest. Thus, we maintain

that our articulatory effort measure represents a reasonable assumption for gauging the net effort spent on the production of a given syllable with increasing duration.

As discussed above, articulation is only half the picture when it comes to effort in speech: effort is also involved in maintaining respiratory energy and vocal fold vibration in speech production, which we subsume under the cover term phonation. We are not aware of any studies on phonatory effort as a function of interval durations in speech. Yet, measures of sub-glottal air pressure during speech have been reported in some studies of speech production, and one may assume that this physiological measure should be correlated with the phonatory effort construct that we propose. Inspection of the measurements presented in these studies (e.g. Ladefoged 1963, Löfqvist et al. 1982) suggests that the effort necessary to sustain phonation during speaking may be hypothesized to be roughly proportional to the duration of time intervals over which phonation is sustained. Thus, we propose the default assumption that phonatory effort be conceived as a simple linear function of syllable duration.

Figure 5.2 plots the hypothetical articulatory (solid gray) and phonatory (dashed gray) effort components as a function of syllable duration, as well as their sum, i.e., overall effort (solid black). It is apparent that the overall effort estimate is a concave increasing function (i.e., an increasing function with decreasing first derivative). It has to be acknowledged that the overall effort in Figure 5.2 only results if the phonatory component is multiplied by an appropriate constant, but we contend that the assumption of overall effort as a concave function of syllable duration is intuitively plausible: lengthening a short syllable entails spending effort on both counts, as the displacement of the involved articulators is increased and additional phonatory effort is spent. At some point, however, articulatory targets are reached, and lengthening the syllable beyond this duration is "costly" only in terms of phonation, but not of articulation.

Despite the separate treatment up to this point, the actual implementation of effort in our model does not explicitly feature separate articulatory and phonatory components. Instead, we have chosen the square root of syllable duration as a rough overall measure that combines both hypothetical sources of effort. While not mathematically equivalent to the solid black trajectory in Figure 5.2, this function fulfills the requirement of concavity, i.e., initial rapid grow, as both articulation and phonation contribute to effort in short syllables, and flattening out with increasing duration, as lengthening syllables beyond the duration at which articulatory targets are fully reached is achieved by sustained phonatory, but not increased articulatory effort. For a given sequence of i syllables with durations s_i, overall effort is defined as a weighted sum of the square roots of individual

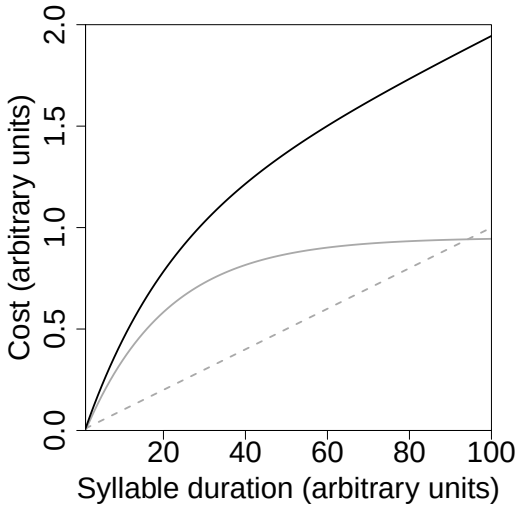


FIGURE 5.2: Plot of the putative estimates of articulatory (solid gray line; negative exponential function with negative base) and phonatory effort (dashed gray line; linear function) as a function of syllable duration, as well as their sum (black line).

syllable durations:

$$E = \sum_i \eta_i \sqrt{s_i}. \tag{5.3}$$

The weights η_i are local trade-off parameters, which adjust the premium placed on production effort for individual syllables, in contrast to the global weighting factor α_E, whose scope is the entire sequence. This implements the assumption that the balance between production and perception requirements may vary also on a local basis, and is implemented for the other cost functions in an equivalent fashion. A possible interpretation of this local parameter will be discussed later. Figure 5.3 shows a plot of cost function E for a hypothetical syllable.

5.2.2 Perception Cost P

Perception cost[3] P in our model implements the hypothesized impetus towards maximizing perceptual clarity, conceptualized as a measure of the difficulty of perceiving a prosodic unit in speech. As our model operates on a syllable basis, this cost function by default operates at the syllabic level. As we said above, however, we are not intending to make any strong claims about syllables as the primary unit of speech perception, and it is conceivable to use any prosodic constituent as the domain of a similar cost function. In this work, we have implemented a perceptual cost function for one further prosodic

[3]We prefer this label over Šimko (2009)'s term *parsing cost*, so as to avoid allusions to high-level linguistic processes (as in "syntactic parsing"), about which our model has nothing to say.

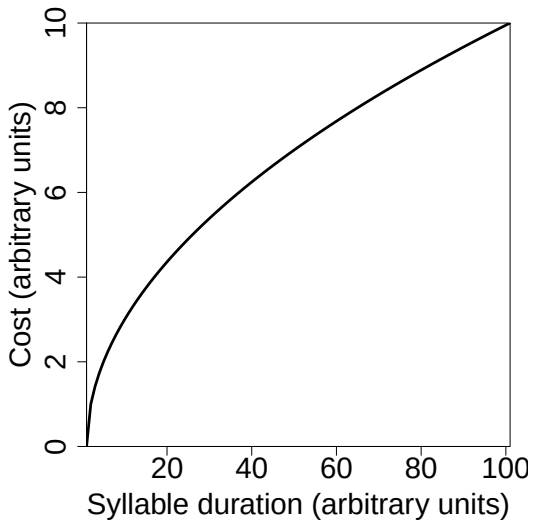


FIGURE 5.3: Plot of cost function E (square root of syllable duration) for a hypothetical syllable.

constituent type, namely the word. In what follows, we will motivate the mathematical concepts used and introduce the components of the perceptual cost function at the syllabic and the word level.

In ETD, the perception-related cost function is implemented as a combination of a spatial and a temporal measure of perceptual clarity. It is clear that in the present model, only the latter dimension can be reasonably evaluated. To recapitulate, this *temporal realization estimate* in ETD is proportional to duration, based on the assumption that longer durations facilitate perception. Crucially, it is non-linear, increasing rapidly up to some threshold, after which it remains virtually unaffected by further increases in duration. The perceptual *cost* associated with duration, then, is the reciprocal of the temporal realization estimate, which is achieved in Šimko (2009)'s model by subtracting it from 1. Figure 5.4, shows the durational component of the perceptual cost in ETD (assuming that the gestural precision estimate, by which it is multiplied, is equal to 1).

The basic assumption that long durations are beneficial for perception is certainly plausible: the articulators are given more time to fully reach their targets, and steady states of vowels can be sustained for longer time, making up for temporary distortions in the acoustic channel or lack of attention on part of the listener. Indeed, empirical evidence reviewed in Chapter 3 indicates that one of the most robust correlates of more listener-centered speech is a decrease in overall speaking rate. Thus, it seems quite reasonable to assume that speakers employ lengthening to secure communicative success. The more interesting question is whether the non-linearity proposed by Šimko (2009) is a realistic assumption. In what follows, we will provide evidence in support of making P non-linear.

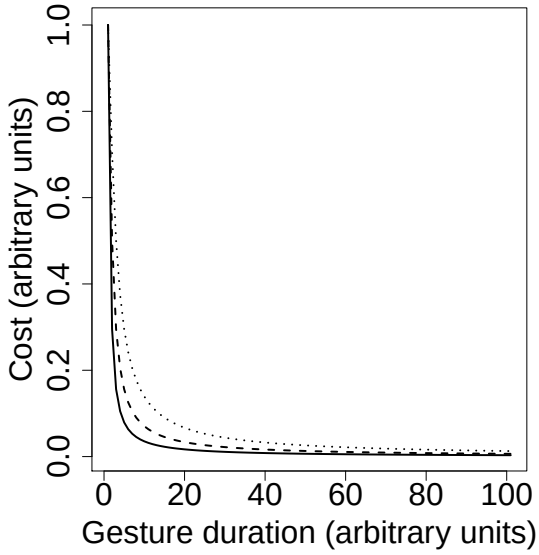


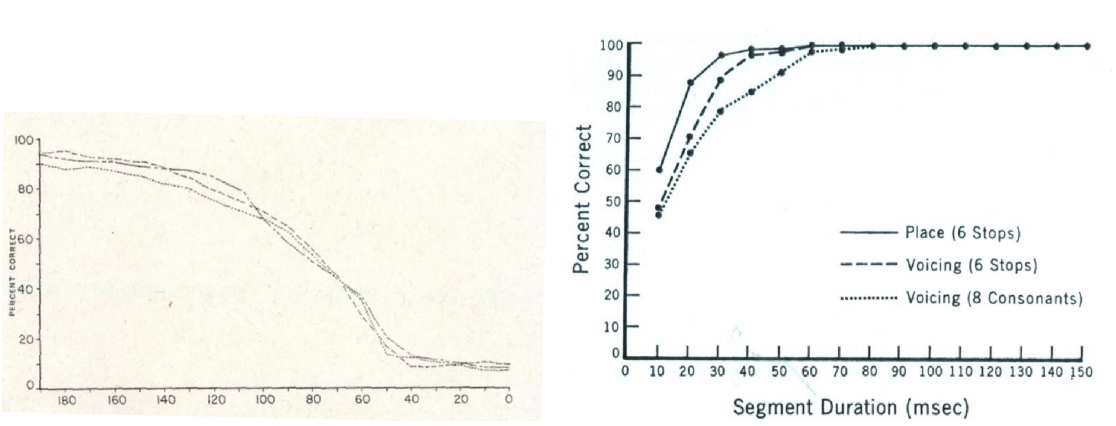
FIGURE 5.4: Plot of the temporal perception cost $(1 - d_g(t))$ in ETD for $c = 2$ (solid), $c = 1$ (dashed) and $c = 0.5$ (dotted).

The hypothesized asymptotic behavior of the perception cost function, to begin with, may be given the following theoretical rationale: the perception cost function can be interpreted as the *inverse of the probability of recognition* of a speech constituent as a function of its duration, i.e., the cost is higher the harder it is to recognize the constituent. One may assume that this probability has an upper bound: at some duration, the recognition rate reaches 100% (or some other upper limit), and nothing will be gained by making the constituent (in our case, a syllable) even longer. This way of reasoning provides a first intuitive motivation for using a function that approaches a lower threshold in the way of Šimko (2009)'s temporal perception cost for P.

Further support is provided by evidence from the psychophysics of duration perception: one prediction that follows from using a convex decaying function of duration as a measure of perceptual difficulty is that progressively longer syllable durations require progressively larger increases in duration in order to achieve the same reduction in perceptual difficulty. This is consistent with common assumptions about *just noticeable differences (JND)* in the duration of acoustic intervals (Lehiste 1970): JND in the acoustic duration between two stimuli are assumed to be proportional to stimulus duration itself, an instance of *Weber's law*, which posits this relationship as a general property of the perception of physical quantities. Research discussed in Lehiste (1970) as well as subsequent studies reviewed by Wagner (2008) (e.g. Friberg and Sundberg 1995, McAuley 1995) suggests that this may be only approximately true for auditory duration judgments, and Lehiste (1970) calls for caution in applying findings on JND of auditory durations, most of which have been obtained using non-speech stimuli, to speech perception. We nonetheless think that it represents a reasonable assumption for

the perceptual cost function in our model. In particular, this line of argument supplies a rationale for employing a *smooth* function – after all, the requirement of a rapid initial decrease followed by a flat trajectory would also be satisfied by simply connecting two straight lines, one with negative and one with zero slope. This, however, would imply that the same increase in duration leads to the same decrease in perceptual cost for a long than for a short syllable, which is at variance with Weber's law.

Finally, there is also more direct empirical evidence that supports the use of a convex decaying function of the kind proposed by Šimko (2009) as a measure of perceptual cost. This evidence comes from studies that employ the *gating* paradigm. In this paradigm, subjects are exposed to acoustic syllable fragments of varying duration and have to indicate the identity of the syllable or of parts of it. Correct recognition scores from this paradigm, provide a measure of perceptual difficulty as a function of syllable fragment duration. Figure 5.5 shows results from two gating studies performed with English speech data and listeners, Grimm (1966) (left; note the inverted x-axis) and Tekieli and Cullinan (1979) (right) on the identification of phonemes and phonetic features from acoustic syllable fragments of varying duration.



(a) (b)

FIGURE 5.5: Recognition scores from gating studies. Panel (a): Recognition scores of English consonant phonemes (pooled over several consonants, for three vocalic contexts) from temporally gated CV syllables as a function of syllable fragment duration (in ms; note the inverted x-axis), reproduced from Grimm (1966). Panel (b): recognition scores of consonantal features from temporally gated English CV syllables as a function of syllable fragment duration, reproduced from Tekieli and Cullinan (1979).

The recognition curves in Figure 5.5 display a remarkable similarity to the temporal perception cost used in Šimko (2009). The perception cost P, under this view, may be imagined as turning these curves "upside down" (which is done in Šimko (2009) by subtracting the temporal realization estimate shown in Figure 5.4 from 1), reducing the cost with increasing recognition rate up to the duration ceiling where recognition

approaches 100%. One important caveat, however, is that stimuli are not shortened in a quite naturalistic way in the gating paradigm: shortening is achieved by simply truncating the signal at a certain point relative to its onset or offset. A better method for the present purpose would be to vary the duration of the steady state of a phone (if applicable), while transitions at the on- and offset are left intact. Still, we would argue that the data presented in Figure 5.5 provide a reasonable approximation of the relationship between speech segment duration and perceptual difficulty, and thus constitute additional evidence to support the usage of a convex decaying function similar to the one used in Šimko (2009) for the perception cost P in our model.

In the course of our work on the model, we have been using two different versions of P. Initially, we have simply employed the reciprocal of syllable duration, $1/s$, for representing the hypothetical perceptual cost. This function differs in one important respect from the one shown in Figure 5.4: it converges towards infinity for very small durations, so that syllables can never be entirely deleted in this version of the model. By contrast, the function used by Šimko (2009) has a finite intercept and thus principally allows for deletions. However, caution is warranted regarding this issue; articulatory studies have raised the question whether there are ever really deletions at the *gestural* level, advancing the alternative hypothesis that apparent deletions in the acoustic domain are consequences of articulatory reorganization, as exemplified by the famous "perfec memory" example from Browman and Goldstein (1990). In this work, we will therefore continue to employ the $1/s$ model as a baseline that has nothing to say about deletion phenomena. We have nevertheless also developed a second version of the model, using a finite intercept function for P similar to Šimko (2009)'s temporal realization estimate, which will be introduced in detail below. We will see that this version of the model does make a very interesting prediction regarding apparent deletions in the acoustic domain, but it exhibits problematic behavior with regard to other phenomena.

From now on, we will refer to the perceptual cost function that applies at the syllabic level as P_S, in order to differentiate it from the word-level perception cost to be introduced below. P_S for a given sequence of syllables is defined as

$$P_S = \sum_i \frac{\psi_i}{s_i}. \tag{5.4}$$

Analogously to E, P_S is also modified by a local weighting factor, ψ_i, which modifies the demands on perceptual clarity for individual syllables. The interpretation of this local parameter is a central aspect of our modeling work: ψ_i constitutes the mechanism by which syllabic prominence is accounted for in the model. In keeping with the concept of prosodic prominence as localized hyperarticulation as discussed in Chapter 3, we assume

that speakers prioritize clarity over effort minimization in prominent constituents, as they are particularly critical for communicative success. Thus, prominence is modeled by locally increasing ψ_i for syllables that are intended to be prominent. This obviously imposes a tendency to lengthen prominent syllables relative to non-prominent ones, in accordance with a large body of empirical findings (e.g. Delattre 1966, Fant and Kruckenberg 1989, Fry 1958, Heuft et al. 2000, Prieto et al. 2012, Streefkerk 2002). As parameter ψ_i applies to individual syllables, we use it to incorporate lexical stress, i.e., the greater prominence of syllables compared to other syllables within the same word. Figure 5.6 plots P_S for $\psi_i = 0.5$ (gray) and $\psi_i = 1$ (black), representing possible values for an unstressed and a stressed syllable, respectively.

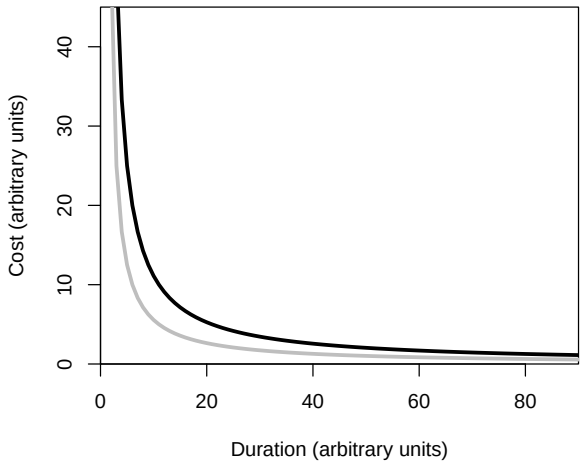


FIGURE 5.6: $P_S = 0.5/s$ ("unstressed", gray) and $P_S = 1/s$ ("stressed", black).

As we said above, while P_S operates at the syllabic level, perception of other prosodic constituents may be modeled in a similar fashion. We have implemented this in this work for one additional constituent type, the word. This choice is motivated mainly by the reasoning that prominence relations apply not only between syllables, but also between words. We hypothesize, based on our considerations in Chapter 3, that these prominence relations express differences in the semantic or pragmatic "weight" between words in an utterance, such that more important words are made more prominent. In particular, we assume that some words in an utterance carry greater semantic or pragmatic "weight" than others and hence receive an *accent*. Based on the findings reviewed in Chapter 3, it is hypothesized that a durational process, accentual lengthening, operates on the duration of the accented word (Cambier-Langeveld and Turk 1999, White 2002). This is accounted for by introducing an additional component of P called P_W. P_W is essentially a copy of P_S, the difference being that it is not a function of individual syllable durations, but of the sum over all syllable durations in a word. Since the model is agnostic towards the propositional content of simulated utterances, we simply define words as arbitrary continuous non-overlapping sub-sequences of S.

P_W is thus defined as

$$P_W = \alpha_{W_j} \sum_j \frac{\Psi_j}{w_j}, \tag{5.5}$$

where w_j stands for word duration, simply defined as the sum over the syllable durations within the j-th word in an utterance.

Parameter Ψ_j, analogously to ψ_i, controls the relative prominence of words. The particular value of α_{W_j} modifies the overall strength of word prominence relative to the other partial cost functions. In our simulations, we will preliminary restrict investigation to the case of one accented word per utterance. This word may be referred to as bearing the *nuclear accent* of the utterance (we will continue to use the term "accent" as a convenient shorthand), corresponding to its *focus exponent* in linguistic terms (Bolinger 1958, Ladd 2008). In the simulations involving accentual lengthening to be reported below, we will therefore simply assume that α_{W_j} is equal to 1, and control the strength of accentual lengthening via the Ψ_j parameter, which assumes a positive value for the accented word, and 0 otherwise. There is no necessity to do do so, and it would certainly be realistic to view word prominence as a gradual parameter (Widera et al. 1997); however, this may not always have durational consequences throughout the word. Indeed, White (2014) hypothesizes that no durational processes affect word duration as a whole in unaccented contexts, and we follow this assumption in our preliminary modeling approach.

The overall perception cost P is the sum of the syllable- and the word-level component, as defined in equation 5.6. There is no mathematical necessity to sum P_S and P_W before entering C, and results would not be different if they entered the overall cost computation as separate terms. Equation 5.6 is employed simply for aesthetic reasons – so as to retain the general form of equation 5.1 – and to demonstrate that P_S and P_W are instances of a general perception mechanism.

$$P = P_S + P_W. \tag{5.6}$$

5.2.3 Duration Cost D

As in ETD, the third component D implements the assumption that speaking rate is to some extent independent of the H&H continuum. D may thus be interpreted as relating to the overall time used for conveying and decoding the sequence as a shared resource between both parties. Since we have no more specific hypotheses about the type of function that should be used for this purpose than Šimko (2009), we simply define D as the weighted sum over all syllable durations in an utterance, with local trade-off weights

δ_i analogous to the other component functions:

$$D = \sum_i \delta_i s_i \tag{5.7}$$

Using a linear function for D predicts the null result that syllable durations are independent of the number of syllables in an utterance, which is in agreement with some empirical results, as reviewed in Chapter 3. By contrast, using non-linear functions for D will make syllable durations dependent on utterance length. Experimentation showed that using convex functions (i.e., functions with increasing first derivative), such as exponentials or powers > 1, predicts shortening of syllables as a function of syllable count in the utterance, whereas using concave functions (functions with decreasing first derivative), such as roots or logarithms, predicts lengthening of syllables as a function of syllable count in the utterance. Since there is no independent evidence for preferring a linear function over these, we can of course not claim that the model's behavior is a result, in the sense of explaining the observed pattern. However, none of the results presented in this thesis hinge on the linearity of D, and we therefore simply use the linear function as a default assumption in the absence of more specific hypotheses.

5.3 Optimization

The model has been coded in the R language (R Core Team 2014). The source code is included in the Appendix of this work. The model uses the implementation of the Nelder-Mead algorithm (Nelder and Mead 1965) in the built-in function *optim* for determining the vector of syllable durations S that minimizes cost function C. The Nelder-Mead algorithm is a *simplex* method for function minimization, which, in a nutshell, can be explained as follows: For an n-variable function to be minimized, the algorithm starts out by arbitrarily selecting $n + 1$ points in the n-dimensional variable space (these points thus form a *simplex*, the simplest possible volume of a given dimensionality). In the present case, n is the length of the simulated utterance, and one point in the variable space corresponds to a given set of n syllable durations for the utterance. The algorithm then computes the function value (in our case the cost C) associated with each point. In the next iteration step, the best of these points, i.e., the one that incurs the lowest function value, is kept and the worst point, incurring the highest function value, is exchanged for another point in the parameter space that is approached by expanding or contracting the simplex, or by mirroring the old point across the centroid of the simplex. The basic principle is that if the new point achieves an improvement over the old ones, search is continued in the direction of the operation that was used to arrive at the new point, whereas in other cases, other directions are tried out. The procedure stops once

some well-defined convergence criterion is reached. In case of successful convergence, the simplex has now contracted around the minimum of the cost function (Brunet 2010).

In contrast to the approach used by Flemming (1997, 2001a) and Katz (2010), this method does not require the computation of derivatives. This may be advantageous, because (as pointed out by Katz 2010) computing derivatives can become very complex for certain types of functions. Apart from this, there is no principled reason for preferring one method over the other. One potential problem for any optimization algorithm mentioned by Nelder and Mead (1965) and Šimko and Cummins (2010), however, is that the algorithm may sometimes converge to non-optimal solutions: this may be the case when the algorithm encounters a *local* optimum in the cost landscape, which does not represent the *globally* optimal solution. As we will be concerned with very simple optimization problems, this issue will probably not be of relevance for the present work. In any case, a simple method was adopted from Šimko and Cummins (2010) in order to safeguard the optimization procedure against such cases: the optimization routine is called iteratively, with small random perturbations of the initial parameters at each new call. The rationale for this procedure is that the random perturbations should help to re-start the search for the globally optimal solution when the optimization "gets stuck" in a non-global optimum.

Figure 5.7 plots the evolution of overall cost of an utterance over the course of two consecutive calls of the optimization routine, amounting to 1000 optimization runs. As can be seen, the cost initially decreases rapidly, but reaches a minimum shortly after 400 optimization runs. The second call of the optimization function, which is marked by the small perturbation in the cost trajectory at about 500 runs, does not result in further reduction in overall cost. Thus one can be quite confident that the globally optimal solution has been found. Plots of this kind are a generally useful diagnostic for checking whether the optimization procedure has converged to a stable and meaningful result – if the cost trajectory were to exhibit a negative slope throughout, this would be a sign that either the problem is too complex and requires more optimization runs to converge, or (more likely, given the generally simple problems we will be concerned with) there is something wrong with the formulation of the problem. The random perturbations method is not an absolute safeguard against falsely selecting a local optimum – it may not work if the gradients around the local minimum are too steep – but as mentioned above, the danger of selecting a local instead of a global minimum is probably not a really vexing one in the present case anyway.

Figure 5.8 shows optimal syllable durations computed by the model for a simulated utterance of eight syllables, with the first, the fourth and the penultimate syllable being stressed and syllables four and five forming an accented word. The following parameters

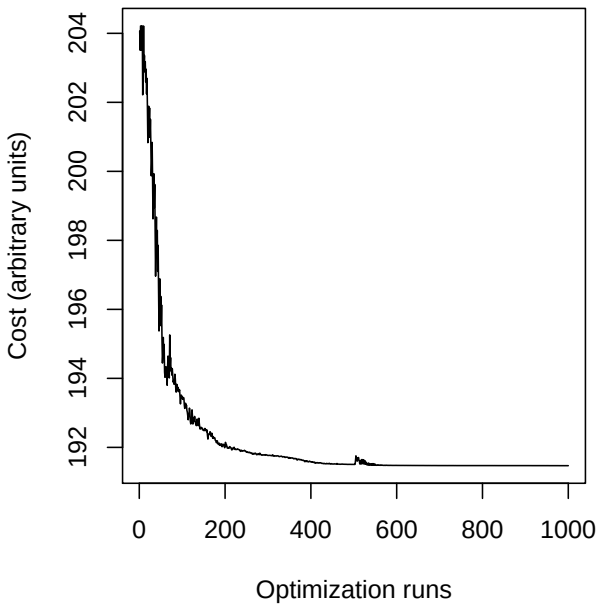


FIGURE 5.7: Plot of the function value of C for a simulated utterance over the course of 1000 optimization runs.

were used: $\psi_i = 1$ for stressed and $\psi_i = 0.5$ for unstressed syllables; $\Psi_j = 2$ for the accented "word", consisting of the fourth and fifth syllable; $\Psi_j = 0$ elsewhere. α_E was set to 3 and all other parameters to 1. Importantly, the exact numerical values of the predicted durations are not of interest, and they do not relate to any real-world units such as milliseconds. As argued in the Introduction to this thesis, the class of models that our approach belongs to is concerned with predicting qualitative patterns rather than exact numerical results, and all theoretically interesting results reported in this thesis will concern duration *ratios*. Hence, predicted durations are reported in terms of arbitrary units.

Inspection of Figure 5.8 suggests that the model captures the basic facts about prominence effects on speech timing in many languages: stressed syllables are longer than unstressed syllables, and all syllables in the accented "word" are lengthened relative to their unaccented stressed counterparts (Cambier-Langeveld and Turk 1999, White 2002). This outcome is of course a direct consequence of the explicit parameter settings, and does not by itself constitute an interesting result beyond demonstrating that the optimization procedure converges. In the following chapters of this work, we will show that the model reproduces a range of durational effects which are not encoded by explicit parameter settings, but emerge as optimal solutions from the interplay of the model's components.

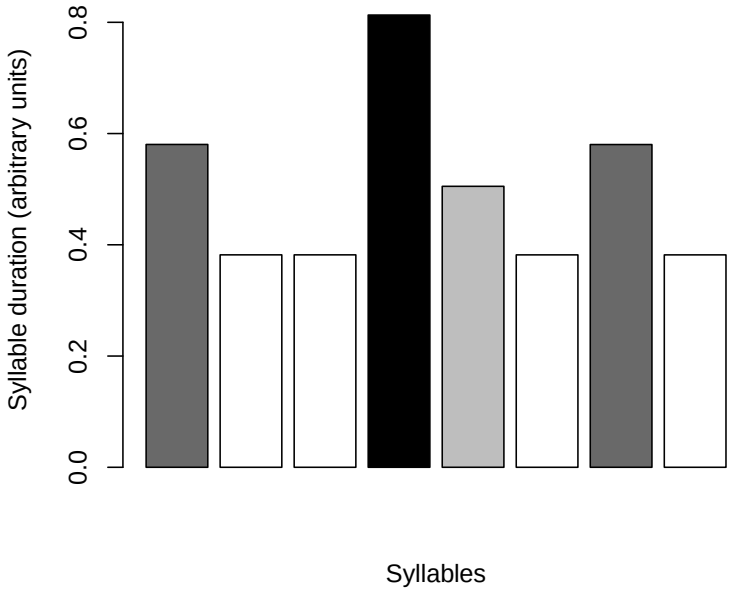


FIGURE 5.8: Syllable durations predicted by the model for a hypothetical utterance with a bisyllabic accented word. Black: +stress +accent; light gray: -stress +accent; dark gray: +stress -accent; white: -stress -accent.

5.4 Discussion

We have introduced the components of our optimization-based model of suprasegmental speech timing, as well as the optimization method itself. The model is very simple, but in our opinion, it fulfills the most important criterion for explanatory models: it is built on well-defined, independently motivated principles. It may of course be debated just how realistic our modeling assumptions are, but we hope to have made our argument reasonably strong, given the high degree of abstraction of our model. Figure 5.9 exemplifies the complete model architecture for a hypothetical utterance.

We have not discussed how position-related lengthening effects are incorporated in the model. The main reason for this is that it is not entirely clear how they are to be interpreted – as we stated in Chapter 3 of this work, we favor the hypothesis that such effects represent actively employed communicative signals over the idea that they are automatic consequences of biomechanical vocal tract properties. This would suggest that positional effects on suprasegmental speech timing effectively represent just another type of prominence; yet, we have seen that both classes of effects behave rather differently with regard to the interactions they are involved in. As we discussed in Chapter 3, the "biomechanical" account cannot be ruled out with certainty based on available evidence, and both accounts may also not be mutually exclusive. We will therefore address the inclusion of position-related lengthening effects in the model in a more explorative fashion, by means of manipulating the local effort parameter η_i. This may be seen as tantamount

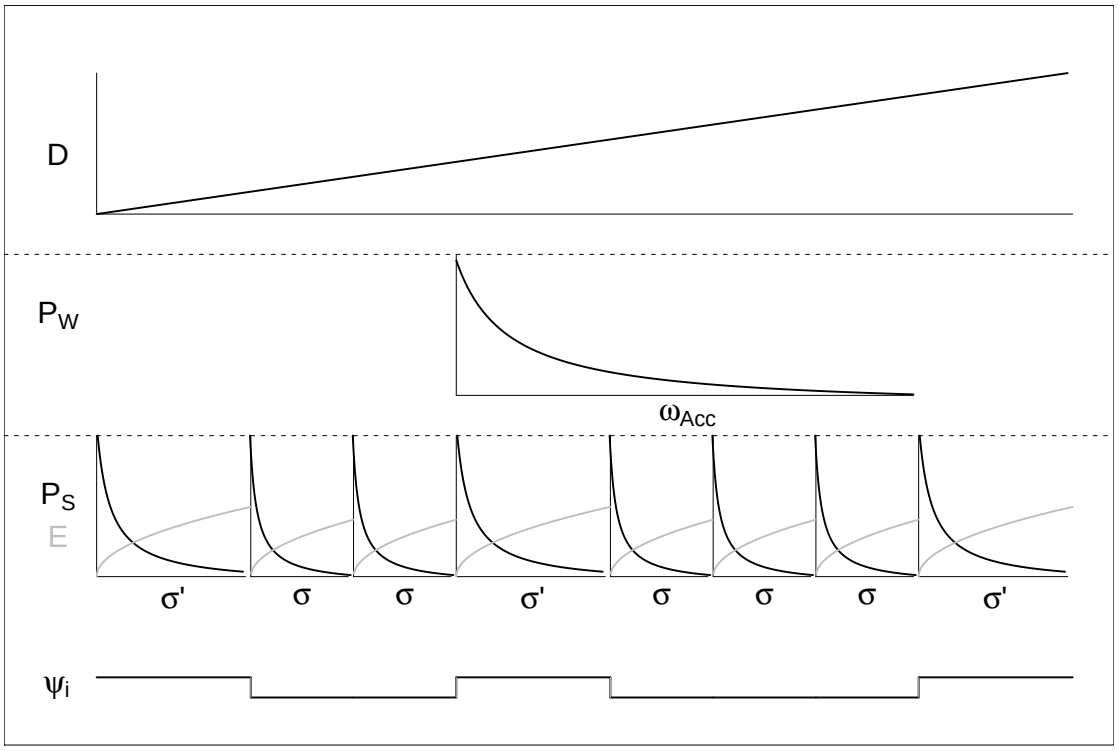


FIGURE 5.9: Model architecture. Cost functions D (utterance level), P_W (word prominence; only shown for accented word W_{Acc}, as parameter Ψ_j is set to 0 elsewhere) and E/P_S (syllabic level; σ; apostrophe denotes stresses) as well as stress parameter ψ_i (other parameters assumed to be constant) are plotted as a function of respective constituent durations for a hypothetical utterance with a tetrasyllabic accented word.

to hypothesizing that position-related lengthening, like prominence, represents a case of local boosting of perceptual requirements, in this case not instantiated by increasing the relative impact of perceptual requirements, but by *decreasing* the relative impact of the effort conservation requirements. This modeling decision would also link positional effects to possible biomechanical causes, although the link is of course rather crude and underspecified. Modeling results related to positional lengthening effect will therefore necessarily have a more speculative character.

It was noted that our model would bear some resemblance to the approach by Flemming (1997, 2001a) and Katz (2010) in the discussion of these works, in that both paradigms are optimization-based and concerned with predicting durational phenomena in the acoustic domain. The major conceptual difference between the two approaches is that our model does not incorporate a concept of "duration targets" that speakers attempt to faithfully reproduce. While we agree with Katz (2010) in asserting that perceptual properties of the speech output to be produced are important in speech production, we are skeptical of the proposition that speakers directly represent duration targets in the acoustic domain as part of their production grammar, and would argue that our production- and perception-related cost functions represent more basic kinds

of processes, not requiring assumptions about high-level representational issues. Indeed, we have seen that this approach (within the embodied optimization paradigm of ETD) is capable of reproducing categorical phonological quantity contrasts, which might appear to constitute a prime example of "duration targets" in speakers' mental representations (Šimko et al. 2014b).[4]

We already briefly touched upon the subject of how our model relates to online processes such as speech planning, and we already answered this question briefly: in the same way as its predecessor ETD, the present model is not to be understood as a real-time production model of speech. While the individual component cost functions are motivated by assumptions about properties of production and perception of concrete speech utterances, we are not endorsing a view of optimization of timing over utterances being computed on-line in speech production. If anything, optimization in our model may be tentatively interpreted with a view to a developmental or evolutionary perspective. On this view, the optimization procedure would be recast as a proxy for a hypothetical process by which interaction among speakers within a linguistic community, or between a child and a caregiver, through repeated exchange converges onto a set of optimal forms that best satisfy the need for successful yet efficient communication. As for the developmental view, we have reviewed one study in which this approach has been fully spelled out using optimization modeling (Howard and Messum 2011). The work by De Boer (2000) has shown how evolutionary development of communicative forms within a community of speakers can be modeled. The optimization procedure in our model, then, may be interpreted as a short-cut for either process. Thus, we would claim that our optimization approach represents a biologically plausible attempt at modeling suprasegmental timing phenomena in speech. In the subsequent chapters of this work, we will examine predictions of the model and evaluate them against the empirically attested patterns we have reviewed in Chapter 3.

[4] A possible way to unify both approaches would be to re-interpret "duration targets" as optimal durations. On this view, a target would be understood not as an abstract duration stored in speakers' mental representations, but rather as something like "a duration that allows comfortable articulation while granting ideal perceptual recoverability of the intended sound". Thus, the model would apply at a higher level, taking the result of optimization for granted. Flemming (1997) seems to go somewhat in this direction, as he discusses the possibility that the formant targets in his first model might themselves be the result of optimization.

Chapter 6

Testing Predictions of Models of Speech Timing

6.1 Introduction

In this chapter, we will investigate two predictions made by some of the models of speech timing reviewed in Chapter 4. As discussed there, several of these models predict that syllable durations shorten as a function of the number of syllables in different larger constituents. The coupled-oscillator model of speech timing in particular, moreover, makes a second strong prediction about speech timing, namely that the duration of the inter-stress interval (ISI) depends on the number of component syllables in different ways, depending on whether the language under investigation is (more or less) "syllable-timed" or "stress-timed". Our model predicts null results for both phenomena; for example, the predicted durations shown in Figure 5.8 do not indicate that stressed or unstressed syllables are lengthened in shorter inter-stress intervals. We will see later that our model does predict a constituent length effect as a consequence of accentual lengthening as proposed by White (2002), but it does not predict ubiquitous constituent length effects in speech, as seem to be implied by some of the models reviewed in Chapter 4.

We have already argued in Chapter 3 that there may be explanations for both durational patterns that do not require the assumption of dedicated timing mechanisms, but as will be explored in more detail below, some open questions remain. These will be subsequently investigated. As for constituent length effects, we will report an empirical investigation on a large speech corpus. The predictions of coupled oscillator models regarding ISI duration will be investigated using simulation experiments on a statistical "toy model" based on minimal assumptions about language-specific timing

patterns and structural properties of different languages, not to be confused with our optimization-based model of speech timing. We will investigate the alternative hypothesis that cross-linguistic differences in the relationship between ISI duration and the number of component syllables are essentially artifacts of language structure.

6.2 A study of Constituent Length Effects in English

6.2.1 Introduction

As we have seen in Chapters 3 and 4, constituent length effects in speech have been widely discussed, and they are a natural prediction of several models of speech timing. The two most thorough empirical studies of constituent length effects, at least in English, yield somewhat converging evidence on the reality of these effects: White and Turk (2010) find large and reliable constituent length effects in accented words, but only a very subtle effect in unaccented environments, restricted to words with initial stress. van Santen (1992)'s corpus analysis shows that vowel duration in accented words varies inversely with the distance (in number of syllables) to the right word boundary, but not with syllable count in the word or the ISI.

These results are compatible with two interpretations. One is that they are indicative of genuine constituent length effect which does not operate at the word or ISI level, but on the interval between the onset of a stressed syllable and the following word boundary. This interval has been termed *word rhyme* White and Turk (2010) or *Narrow Rhythm Unit* (NRU) Bouzon and Hirst (2004), Jassem (1952), and Jassem (1952) explicitly proposed it as the domain of a temporal equalization process in English. According to this theory, speakers of English attempt at regularizing NRU duration, which would predict a constituent length effect in this unit. Syllables not contained in an NRU, i.e., unstressed syllables occurring before the main stress of the word they are part of and thus within the so-called *anacrusis*, are produced as rapidly as possible according to Jassem (1952)'s theory. On this account, models of speech timing that predict constituent length effects could be "rescued" by positing the NRU as the domain of the effect.

An alternative interpretation, suggested by White and Turk (2010), is that the observed pattern is the result of a progressive word-final lengthening effect: vowels are longest when they are directly adjacent to a word boundary, and become shorter with added intervening syllables. For stressed vowels, both hypotheses are indistinguishable – the number of syllables between the stressed syllable onset and the right word edge is, by definition, the same as the syllable count in the NRU. For unstressed vowels, however, it is possible to pit the number of syllables in the NRU against the number of syllables

to the right word boundary: the NRU compression hypothesis predicts the vowel in the word-final syllable to be shorter in "Minister" (trisyllabic NRU) than in "Mister" (bisyllabic NRU), whereas the progressive word-final lengthening hypothesis predicts no such difference.

Results of a corpus analysis by Hirst (2009) appear to favor the word-final lengthening hypothesis: the constituent length effect at the NRU observed in earlier studies on the same data Bouzon and Hirst (2004), Hirst and Bouzon (2005) does no longer hold if NRU-initial, -final and -medial phones are analyzed separately. However, this study did not control for phrasal prominence, and, additionally, conflated consonants and vowels. Since the final phone of an NRU will presumably often be a coda consonant, it is not clear how to interpret the result of the study, given that White and Turk (2010) found only nuclei, not coda consonants to shorten with syllable count in the NRU (or distance to the right word boundary). We will report a reanalysis of the same data, in order to assess possible influences of NRU length on vowel duration whilst controlling for prominence and positional factors. We will also investigate possible effects of syllable count in the word and in the ISI, in order to provide a replication of the studies by White and Turk (2010) and van Santen (1992) on somewhat more naturalistic data.

6.2.2 Corpus Analysis

6.2.2.1 Material and Methods

Analyses were conducted on speech data from the Aix-MARSEC corpus (Auran et al. 2004). This corpus comprises approximately $5^1/_2$ hours of automatically segmented and prosodically transcribed broadcast speech, produced by 17 male and 36 female speakers of British English. Analyses were carried out on vowel durations, using the existing segmentation of the data. A number of measures were taken in order to avoid confounding of results. Vowels from utterance-initial and final words were excluded from the analysis, so as to avoid potential effects of initial and final lengthening. Moreover, we discarded data from a number of words in the corpus for which stress was marked on more than one syllable, as it is not clear how to define units such as the ISI in such cases. Analyses were carried out on z-normalized vowel durations, i.e., the duration of each vowel token is represented by how many standard deviations it differs from the mean duration of all tokens of the respective phoneme, and this mean value is the subtracted, so as to center the duration distribution around zero. This normalization method eliminates inherent vowel duration differences as a potential confound.

Finally, we controlled for two prosodic variables: first, a variable termed PROMINENCE was defined using the existing prosodic transcription of the corpus, comprising three

prominence levels, stressed accented (1), stressed unaccented (2) and unstressed (3). These will be referred to as *S +Acc*, *S -Acc* and *U*, respectively. Second, we identified word-final vowels and defined a control variable WITHIN-WORD-POSITION with the levels *final* and *non-final*. Vowels from monosyllabic words were counted as word-final. We chose not to control for any other variables, such as the phonological environment of a vowel, syllable type, or between-speaker variation. In doing so, we assume that potential confounding variables not accounted for should be randomly distributed with respect to the experimental variables of interest, or that they should cancel each other out to some extent. We thus incur some risk of factor confounding, but the alternative would be to accept substantially reduced cell sizes, which by itself would make results less robust. As we shall see, our control methods are rigorous enough to yield consistent results, which are in agreement with findings from more controlled studies.

Vowel durations were analyzed using quantile regression, as implemented in the R package *quantreg* Koenker (2005). Quantile regression allows for computing median estimates, which arguably yields a more accurate representation of vowel durations than techniques that provide mean estimates, as vowel duration distributions typically exhibit a considerable positive skew. We applied a stepwise analysis procedure: first, we fitted a model with the factors PROMINENCE (*S +Acc/S -Acc/U*) and WITHIN-WORD POSITION (*final/non-final*) to the data. This model will be referred to as *basicmodel*. We then created a dummy variable, referred to as CONTROL, which comprised all combinations of factor levels of PROMINENCE and WITHIN-WORD POSITION. In a second analysis step, we constructed three separate regression models, one for each of the three constituent types, word, ISI, and NRU. In each of these models, slopes for vowel duration by syllable count in the respective constituent type were nested within the levels of CONTROL, using R's "/" operator (Chambers and Hastie 1992). Thus, constituent length effects were tested separately within the subsets of the corpus defined by the combinations of PROMINENCE and WITHIN-WORD POSITION, so that confounding by these factors was eliminated.

The three models used for testing the individual constituent types will be referred to as *wordmodel*, *isimodel*, and *nrumodel*. Data from cells as defined in these models that contained less than 100 observations were discarded. After all exclusions, approximately 40000 vowels remained to be analyzed in *wordmodel* and *isimodel*. For *nrumodel*, there were additional exclusions, as detailed below.[1]

[1] in Windmann et al. (2015a), we applied Bonferroni correction in these analyses. Following suggestions such as Rothman (1990) and Perneger (1998), we now think that this is unnecessary, and maintain that theoretically informed interpretation of results is a sufficient guard against false positives. We will report all effects with $\alpha = 0.05$ here, but the interpretation of results will be the same as in the Windmann et al. (2015a) paper.

6.2.2.2 Results

We will begin by discussing the results of *basicmodel*. Planned comparisons showed non-word-final *S +Acc* vowels to be significantly longer ($t = 3.93; p < 0.0001$) and *S -Acc* vowels to be significantly *shorter* ($t = -7.84; p < 0.0001$) than *U* vowels in the same position. In word-final position, *S +Acc* vowels are also longer than *U* vowels ($t = 17.84; p < 0.0001$); the difference between word-final *S -Acc* and word-final *U* vowels is also significant ($t = 2.40; p < 0.05$). *U* vowels are longer in word-final than non-word-final position ($t = 14.67; p < 0.0001$), as are *S +Acc* ($t = 19.20; p < 0.0001$) and *S -Acc* ($t = 15.10; p < 0.0001$) vowels. There is also evidence for an interaction: the durational difference between *S +Acc* and *U* vowels is greater ($t = 8.66; p < 0.0001$) in word-final than in non-word-final position. The durational difference between *S -Acc* and *U* vowels is smaller in non-word-final than in word-final position ($t = -2.58; p < 0.01$). Thus, there are reliable effects of prominence and within-word position on vowel duration and an interaction between both. Results are graphed in Figure 6.1.

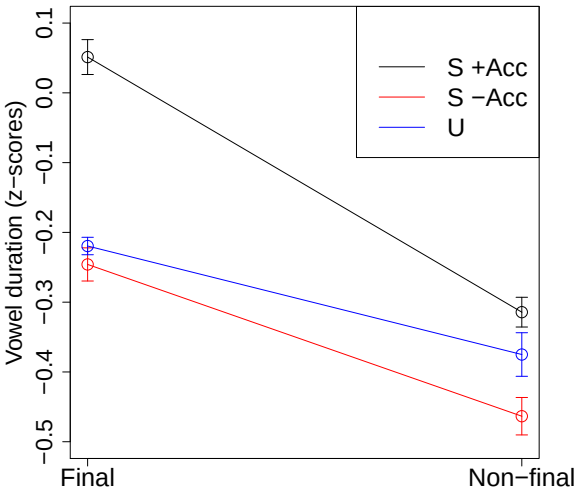


FIGURE 6.1: Z-normalized vowel duration (medians and 95% confidence intervals) by prominence and within-word position in the Aix-MARSEC corpus.

The surprising finding of greater *U* than *S -Acc* vowel durations is most likely an artifact of the z-score normalization: as is well-known, the distribution of English vowel phonemes in stressed and unstressed syllables is near-complementary; most vowel phonemes in the corpus appear almost exclusively either in stressed or in unstressed syllables (with the exception of short high vowels, for which stress indeed makes little difference). Since the z-score normalization sets the mean durations of all vowel phonemes to zero, duration differences between stressed and unstressed vowels are largely eliminated. The category of *S -Acc* vowels in particular comprises mainly those observations from the lower tail of the stressed vowel duration distribution, so that *S -Acc* vowels appear

to be even shorter than *U* vowels. Lexically stressed and unstressed vowel durations are thus not directly comparable using our data and method.

Figure 6.2 graphs z-normalized vowel durations (medians and 95% confidence intervals) by syllable count in the ISI (the interval between two consecutive stressed syllable onsets), for all data (left panel), and separately for word-final (middle panel) and non word-final vowels (right panel). The different colors denote the three levels of prominence. The individual trajectories in the middle and right panel correspond to the nested slopes in *isimodel*. This way of presenting the data highlights the benefits of our nested analysis: as long as the data are pooled across within-word positions, there seem to be clear effects of syllable count in the ISI, especially in S +Acc vowels. Once within-word position is controlled, a different picture emerges: for word-final vowels, there is some evidence compatible with a constituent length effect at the ISI level in *U* and *S-Acc* vowels, which is corroborated by *isimodel* yielding significant negative slopes for syllable count in the ISI in word-final *U* ($t = -5.04; p < 0.0001$) and in word-final *S -Acc* vowels ($t = -3.15; p < 0.05$; note that these effects may be underestimated by the slopes of our model, which assume linear effects of duration by syllable count). None of the remaining nested slopes are significant.

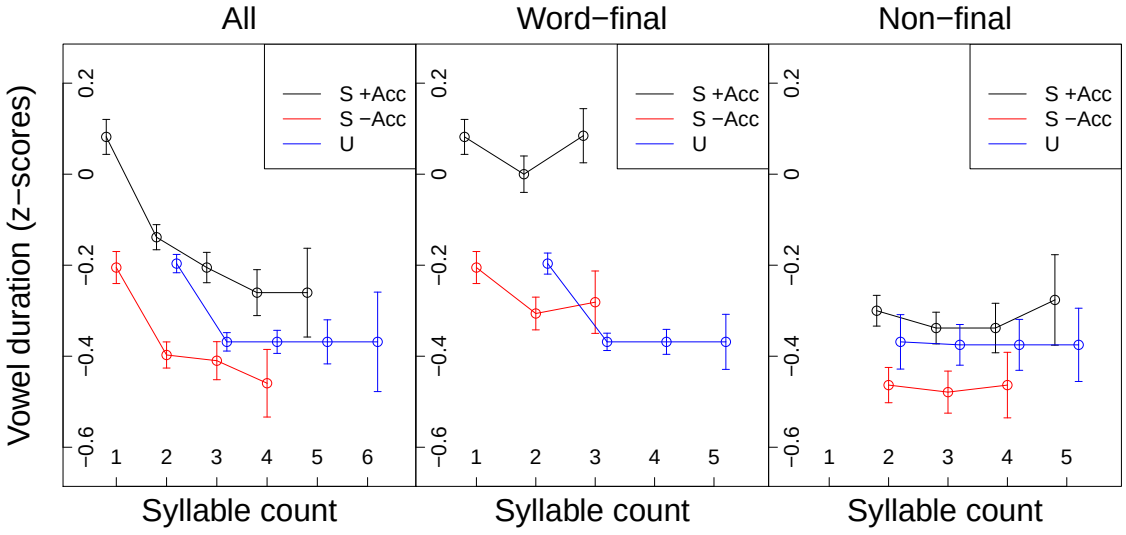


FIGURE 6.2: Z-normalized vowel duration by prominence level, within-word position and syllable count in the ISI in the Aix-MARSEC corpus.

Figure 6.3 graphs z-normalized vowel durations from the MARSEC corpus as a function of the number of syllables in the *word*, in the same fashion as Figure 6.2 above. The pattern of results is the same as in the ISI analysis: as long as within-word position is not controlled, there seems to be a shortening effect of the number of syllables in the word on vowel duration, particularly for *S +Acc* vowels. Once within-word position

is controlled, the effect of word length on vowel durations turns out to be essentially random. There is marginally significant shortening of vowel duration as a function of word length in word-final *S -Acc* vowels ($t = -1.81; p = 0.07$) and weak effects in the direction of *lengthening* in word-final *S +Acc* ($t = 2.10; p < 0.05$) and non-word-final *S -Acc* vowels ($t = 2.10; p < 0.05$). The inconsistent overall pattern suggests that these effects are most likely spurious. One may note, however, that there is a quite distinct durational pattern in word-final *U* vowels. We will provide a possible interpretation of this pattern below.

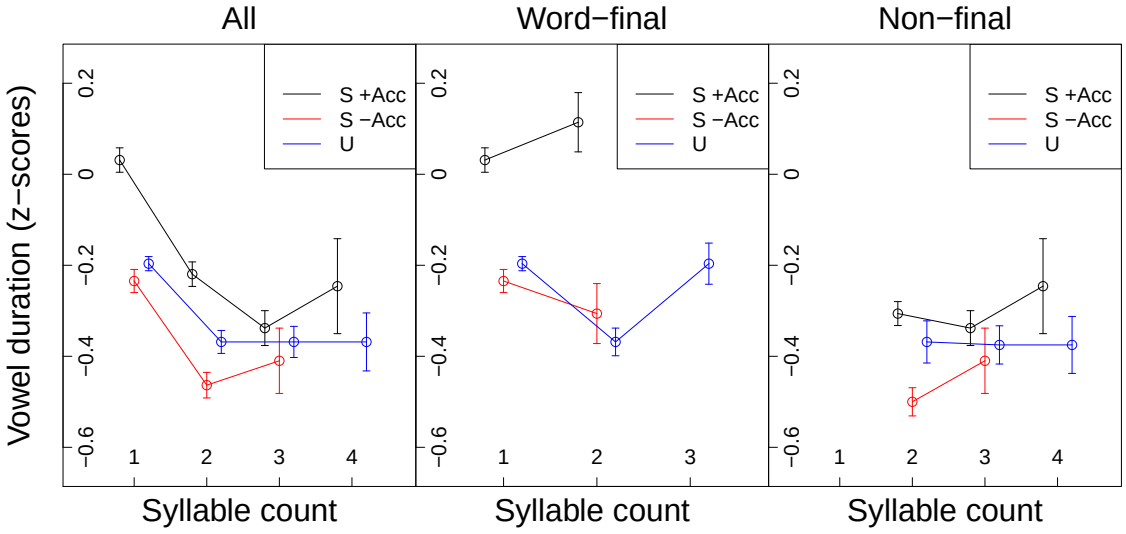


FIGURE 6.3: Z-normalized vowel duration by prominence level, within-word position and syllable count in the word in the Aix-MARSEC corpus.

Figure 6.4 graphs z-normalized vowel durations from the MARSEC corpus as a function of the number of syllables in the *NRU* (the interval between the onset of a stressed syllable and the following word boundary), in the same fashion as in Figures 6.2 and 6.3 above. For the NRU analysis of unstressed vowel duration, we excluded observations from syllables in *anacruses*, i.e., unstressed syllables occurring before the stressed syllable within a word (or within words that do not contain a stressed syllable at all), as these are not part of the NRU according to Jassem (1952)'s model. Since syllable count in the NRU is, by definition, one for word-final stressed vowels, these were also excluded from *nrumodel*, but are shown in Figure 6.4. These exclusions lead to approximately 19000 observations being included in *nrumodel*.

Inspection of Figure 6.4 tentatively suggests a progressive word-final lengthening effect in *S +Acc* vowels, in accordance with results by van Santen (1992) (recall that for stressed vowels, syllable count in the NRU is isomorphic to the number of syllables between the vowel and the right word boundary). The difference between *S +Acc* vowels

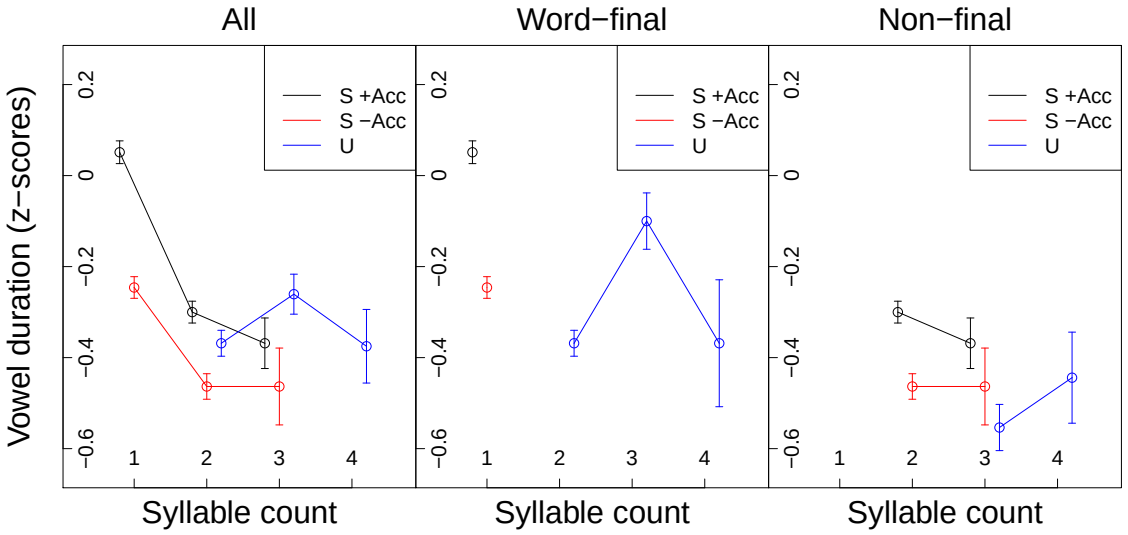


FIGURE 6.4: Z-normalized vowel duration by prominence level, within-word position and syllable count in the NRU in the Aix-MARSEC corpus.

from bi- and trisyllabic NRU, and, hence, between penultimate and antepenultimate *S +Acc* vowels, however, is only marginally significant ($t = -1.93; p = 0.054$). Crucially, there are no effects compatible with a constituent length effect at the NRU level in unstressed vowels. The nested slope for word-final *U* vowels suggests a *lengthening* effect ($t = 10.56; p < 0.0001$), but graphical presentation of results in the middle panel of Figure 6.4 indicates a more complex pattern, similar to the result observed in the word-level analysis.

Figure 6.5, finally, clarifies why constituent length effects are observed in uncontrolled data: shown are the percentage of word-final vowels as a function of constituent length in the MARSEC corpus: for example, 100% of all stressed vowels in monosyllabic ISI come from word-final syllables, which is not surprising, given that a monosyllabic ISI is defined as a primary stressed syllable followed by another primary stressed syllable, so that there is necessarily a word boundary intervening. In bisyllabic ISI, this proportion is only about 60% for stressed syllables, and it decreases further with increasing ISI length. The resulting trajectories bear a striking resemblance to the durational results obtained without controlling for within-word position, particularly for stressed vowels. As for unstressed vowels, results are not obviously related to the proportion of word-final observations, and we will argue below that there is another factor that needs to be taken into account. One interesting observation with regard to unstressed vowels, however, can be made in the left panel of Figure 6.5: while the proportion of word-final observations does decrease as a function of ISI length for unstressed vowels, the effect is much weaker than for stressed vowels. A possible explanation is that longer ISI are

likely to involve polysyllabic words. In this case, some of the unstressed syllables in a long ISI will be word-initial or medial, resulting in a weaker correlation between ISI length and the probability to occur word-finally for unstressed syllables. This provides a possible explanation for Kim and Cole (2005)'s negative finding on ISI-based shortening in unstressed vowels.

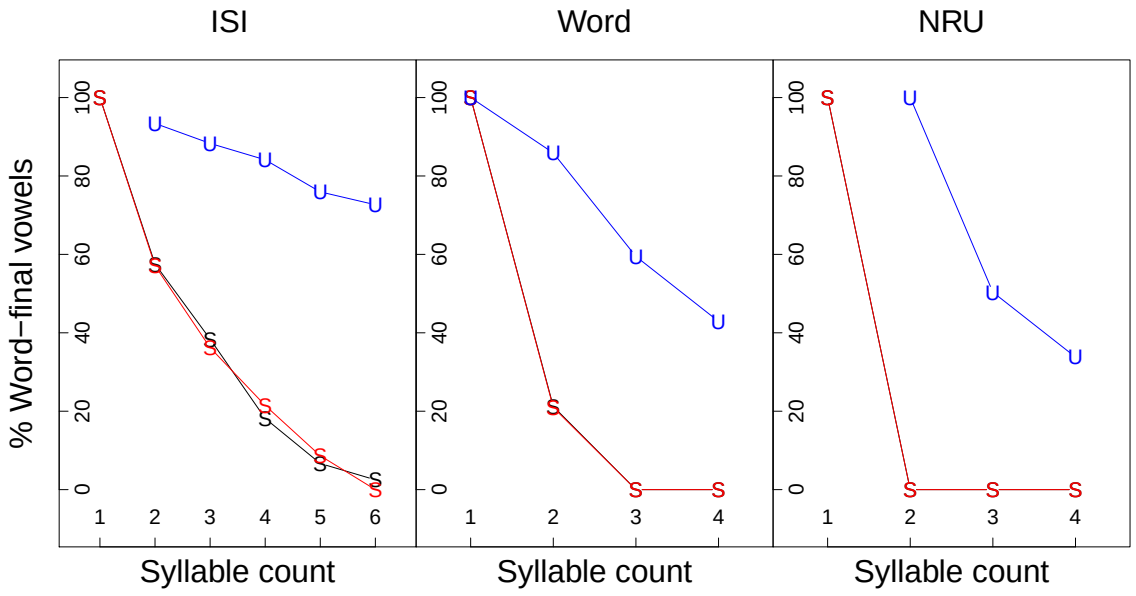


FIGURE 6.5: Percentage of word-final vowel observations as a function of the number of syllables in the three constituents in the Aix-MARSEC corpus, separated by levels of prominence (black: S +Acc; red: S -Acc; blue: U).

6.2.3 Discussion

To summarize, results of the corpus analysis do not support shortening effects at the level of any of the constituents investigated. Such effects seem to be pervasive if within-word position is not controlled. Once position within the word is accounted for, apparent constituent length effects are no longer observed. This strongly suggests that constituent length effects reported in earlier studies (Bouzon and Hirst 2004, Campbell 1988, Kim and Cole 2005, Williams and Hiller 1994) are an artifact of word-final lengthening.

The data do provide evidence for two localized lengthening effects, accentual and word-final lengthening, lending support to White (2002)'s domain-and-locus approach towards speech timing. As argued above, the effect of lexical stress cannot be assessed in the MARSEC corpus, due to the complementary distribution of vowel phonemes in stressed and unstressed syllables. Word-final lengthening seems to be pervasive, and our data also suggest an interaction of word-final lengthening and accent. One caveat, however,

is that our data do not definitively establish the word as the trigger of this lengthening effect – it may be the case that the lengthening of word-final vowels is really instantiated by some intermediate prosodic phrase that has simply not been marked in the corpus annotation. This may be a distinct possibility, given that our review in Chapter 3 has yielded mixed results for word-final lengthening. Yet, the general conclusion remains that apparent constituent length effects in our data are an artifact of such localized lengthening phenomena.

The analysis revealed two durational patterns that are not accounted for by prominence or final lengthening: first, word-final *S -Acc* vowels are longer in monosyllabic than in bisyllabic ISI, and word-final *U* vowels are longer in bi- than in trisyllabic ISI. Similar patterns have been observed in earlier experimental studies (e.g. Fowler 1977, Lehiste 1972). While these findings are in the direction of a constituent length effect, this may not be the preferable explanation – no comparable tendency is observed in non-word-final vowels, and in either case, the difference only resides in the comparison between vowels from ISI with minimum versus larger syllable count. A unified explanation may be suggested based on the fact that the difference in either case is whether the critical syllable is followed by a stressed or an unstressed syllable across the word boundary: in the case of a final stressed syllable, mono- and bisyllabic ISI correspond to $S\#S$ and $S\#U$ sequences, respectively ('#' denoting the word boundary). For an unstressed final syllable, bi- and trisyllabic ISI correspond to $SU\#S$ vs. $SU\#US$ sequences. Thus, the durational effect can be interpreted as "stress-adjacent lengthening", following White (2002)' reanalysis of similar findings. White views this phenomenon as a rhythmic effect that is employed in order to create durational contrast, reminiscent of the widely discussed *stress clash* phenomenon.

Second, NRU length seems to have an alternating effect on unstressed vowel duration: for word-final vowels, the pattern could be described as $2 = 4 < 3$ (where the numbers represent syllable count in the NRU), whereas there seems to be a $3 < 4$ pattern. This finding may be straightforwardly explained as a secondary stress effect: a word-final unstressed syllable in a trisyllabic NRU is one unstressed syllable removed from the preceding stress ($SU[U]\#$). The assumption of secondary stress assignment would account for the greater duration of vowels in this position relative to word-final vowels in bi- and tetrasyllabic NRU. For non-final unstressed vowels, the situation is reversed: in the case of a trisyllabic NRU, this vowel comes from the syllable directly adjacent to the stressed syllable ($S[U]U\#$), whereas in the tetrasyllabic case, the non-final unstressed category includes observations from the unstressed syllable that is one syllable removed from the initial stressed one ($S[U\mathbf{U}]U\#$), which is a potential site for the putative secondary stress effect. This is consistent with the $3 < 4$ pattern for unstressed vowels in the right panel of Figure 6.4. The analyses by ISI and word largely mask this effect, but

it is visible in the word-final unstressed data by syllable count in the word in the middle panel of Figure 6.3. As an explanation of this effect, one may invoke the assumption that *eurhythmic* principles play a role in speech production (Wagner 2002): according to this assumption, languages prefer alternating strong-weak patterns and penalize sequences of prosodically weak syllables. The explanation thus recurs to a similar mechanism than the "stress-adjacent lengthening" account discussed above, although the effect is in the opposite direction.

One final point in which our results disagree with those of White (2002) is that we find no evidence for a constituent length effect at the word level in S +Acc vowels. One may also note that we have not distinguished unstressed vowels from accented and unaccented words. This had initially been done, but we found no durational differences, and collapsed both categories. We certainly do not want to imply that the patterns reported in White (2002) and other rigorous experimental studies are invalidated by our rather coarse analysis. One potential reason for the apparent discrepancy may have to do with the definition of accents in the MARSEC corpus: the documentation of the corpus data in Bouzon (2004) suggests that accent labels in the MARSEC corpus simply refer to any salient tonal movement. It may be speculated that effects such as those discussed by White (2002), i.e., lengthening of all constituents of an accented word and shortening of the individual syllables in longer words, may be restricted to very prominent, possibly nuclear and contrastive accents. Thus, such effects may not be visible in our data. Contrary to White (2002), we have also not found evidence for progressive word-final lengthening in S -Acc vowels. Since the effect in White (2002)'s data was very small, it may be the case that it is simply not detectable in our data due to the higher amount of noise, or that it depends on speaking style characteristics.

Our results are problematic for some of the models of speech timing reviewed in Chapter 4, which predict that constituent length should be ubiquitously observable in speech. The analysis reproduces findings for the word and ISI level from the studies by van Santen (1992) and White (2002) and extends them by showing that the NRU is also not likely to trigger an effect of this kind in English. Earlier findings that appear to support constituent length effects, particularly at the ISI level, are due to failure to control for word-final lengthening. Our conclusions are of course valid only for English and will hopefully stimulate cross-linguistic comparisons in future work.

These results preliminarily suggest that it is unnecessary to invoke a dedicated mechanism to account for constituent length effects throughout speech utterances in an explanatory model of speech timing. Figure 6.6 graphs results of a simulation of a subset of data from the MARSEC corpus, conducted with our optimization-based model. The model was run on input sequences that were based on 2000 actual utterances from the

corpus in terms of number of syllables and locations of lexical stress and word boundaries. The same parameter settings were used as in the example simulation shown in Figure 7.3.1, with two exceptions: α_{PW} was set to 0 throughout, and η_i was set to 0.5 for word-final syllables, in order to simulate word-final lengthening. Unsurprisingly, the simulation captures the apparent constituent length effect at the ISI level found in the real data when within-word position is not controlled, due to the correlation between ISI length and the proportion of word-final syllables. For this subset of the corpus data, there is at best a very weak effect of ISI length on unstressed syllable duration, due to the weaker correlation with the proportion of word-final syllables, as explained above. This suggests an explanation for results reported by Kim and Cole (2005) and shows that it is not necessary to invoke an oscillatory mechanism including additional ad-hoc assumptions to account for this pattern of results. The inclusion of word-final lengthening is sufficient. We are of course not claiming that our model offers an explanation of this pattern of results – rather, the structure of the data itself accounts for the durational pattern. As we have argued, however, our results do not invalidate experimental findings on polysyllabic shortening in words with nuclear or contrastive accent, and we will continue to investigate this phenomenon within our modeling paradigm below.

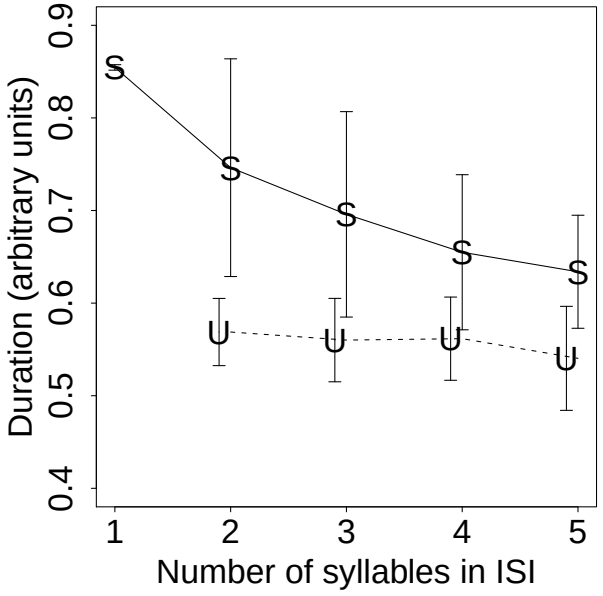


FIGURE 6.6: Simulation of MARSEC corpus data. Shown are simulated stressed (S) and unstressed (U) syllable durations as a function of syllable count in the ISI. No explicit ISI level, but only word-final lengthening is modeled (see text for details).

Results of the corpus analysis may be put in a broader theoretical context by recurring to the discussion of *periodic* versus *contrastive* speech rhythm in White (2014). Our results do not support the former concept, which relates to precisely the kind of temporal periodicities implied by oscillatory models, but they do support the notion

of contrastive rhythm, i.e., the assumption that speech is characterized by alternating strong-weak patterns. This is particularly true for the stress-adjacent lengthening and the secondary stress effect we observed. It would of course be desirable for our model to provide a unified explanation for both phenomena, but we currently see no way of incorporating a truly explanatory account of these phenomena within our rather simple modeling paradigm. One might implement an additional component cost function that instantiates a tendency towards alternation in syllable durations, but this would be little more than a description of the observed facts and not really constitute independent motivation. We hypothesize that a satisfactory explanatory account of the alternation patterns we observed would require more in-depth modeling of word recognition (Cutler and McQueen 2014 and references therein), probably including explicit assumptions about concepts such as attention modulation in listeners. We will not attempt at providing such an integrated account in this work, and leave this enterprise for future research.

6.3 Investigating Regression Results on Inter-Stress Interval Duration

6.3.1 Introduction

The previous section confirmed and extended findings from the literature, suggesting that in English, constituent length effects observed throughout utterances, as predicted for example by coupled oscillator models of speech timing, are spurious. In view of this outcome, on may wonder whether there are similar alternative explanations for the second key prediction of oscillatory models, language-specific differences in ISI duration expressed as a function of the number of component syllables. To recapitulate, Eriksson (1991) found intercept values clustering around 200 ms for "stress-timed" and 100 ms for "syllable-timed" languages in this type of analysis, and O'Dell and Nieminen (1999) showed that this difference is borne out by the coupled oscillator model in an elegant fashion if asymmetrical coupling between a syllabic and an ISI oscillator is assumed. In the discussion in Chapter 4, we brought up differences in stressed/unstressed duration ratios as a possible alternative explanation. However, results by O'Dell and Nieminen (2001), who found a substantial positive regression intercept despite minimal stressed/unstressed duration ratio in an analysis of this type in Finnish, show that there may be yet other factors at play in accounting for the cross-linguistic differences observed by Eriksson (1991). In this section, we will test the hypothesis that these differences, too, may be related to the distribution of stressed syllables relative to word boundaries. This will be done using simulated data generated with a statistical "toy

model" – not to be confused with the optimization-based model introduced in the previous chapter – that determines syllable durations based on minimal assumptions about language-specific timing patterns and structural properties of different languages. This model is of course highly simplistic, as it leaves many other potentially relevant factors unaccounted for. Our investigation therefore should be viewed as a proof-of-concept study whose aim is to highlight a potential alternative explanation for observed facts, rather than to establish definitive conclusions. Our analysis will concentrate on English and two other languages that will be shown to represent extreme cases regarding the alignment of stressed syllables and word boundaries: Finnish and French[2]

6.3.2 Analysis

As we are not aware of regression analyses of ISI duration by syllable count in French and Eriksson (1991)'s analyses of English are based on a rather small data set (actually on average values reported in Dauer 1983), we will first supply an analysis of this kind on a large corpus of French and also report an analysis on the MARSEC corpus for English. As for French, data from the C-PROM corpus (Avanzi et al. 2010) were analyzed. This corpus was compiled for the study of prosodic prominence in French and contains binary syllabic prominence labels from two expert annotators. It comprises approximately one hour of mostly read speech from different genres produced by 16 male and 12 female speakers of French. The C-PROM corpus is thus comparable to the MARSEC corpus, although considerably smaller. ISI boundaries were determined using the existing prosodic transcriptions of the corpora. Following Fant et al. (1991a) and Wenk and Wioland (1982), the ISI in French, in contrast to English was defined as the interval between the *offset* of a stressed syllable and the offset of the following one, as French prosodic structure is assumed to be organized into accentual phrases with final prominence. Linear regression models were fitted to ISI duration by number of component syllables. Only data from utterance-medial ISI were analyzed. Table 6.1 summarizes the corpus data and Table 6.2 summarizes the regression models, including the values from O'Dell and Nieminen (2001)'s analysis of Finnish.

TABLE 6.1: *Number of syllables by syllable count in the ISI in the Aix-MARSEC corpus (English) and the C-PROM corpus (French).*

Language	1-ISI	2-ISI	3-ISI	4-ISI	5-ISI	6-ISI	7-ISI
English	4321	6622	3987	1370	354	46	7
French	222	322	333	237	189	97	79

[2]We will conveniently refer to prominent syllables in French as "stressed", even though this is technically incorrect, as prominence assignment is supposed to be post-lexical in French (Hyman 2006).

TABLE 6.2: *Regression models of ISI duration by number of component syllables in English, French and Finnish. The column headed "r" denotes the "coupling strength ratio", intercept divided by slope.*

Language	Intercept (ms)	Slope (ms)	r	R^2
English	143	123	1.16	0.49
French	122	142	0.85	0.78
Finnish (O'Dell and Nieminen 2001)	104	145	0.71	0.73

The coefficients of the regression model fitted to the English data do not quite match the values reported by Eriksson (1991), but they are in line with the prediction of the coupled oscillator model, in that the "coupling strength ratio" r between intercept and slope does lie in the "stress-timed" region > 1. It is also substantially higher than the equivalent value computed for French, which would be classified as syllable-timed according to this analysis. O'Dell and Nieminen (2001)'s Finnish data seem to pattern more or less with our French analysis. The English and French data thus appear to support the predictions of the coupled oscillator model, assuming a dominant ISI oscillator in English and a dominant syllabic oscillator in French. For Finnish, the situation is less clear-cut; Finnish has been described as *mora-timed*, and we are not aware of explicit predictions made by oscillatory models for this putative class of languages. In any case, the more interesting question regarding the Finnish data his how the substantial positive regression intercept can be borne out regardless of minimal stressed-unstressed syllable duration differences. In what follows, we will suggest differences in the alignment of stressed syllables and word boundaries as a unified explanation for the pattern of results shown in Table 6.2.

We have seen for English that, as syllable count in the ISI increases, the proportion of word-final syllables decreases, particularly for stressed syllables, as shown in Figure 6.5 above. This pattern may differ cross-linguistically, based on the stress-assignment rules of individual languages. French and Finnish represent two extreme cases in this respect. In French, with its supposedly post-lexical assignment of prominence to the final syllable in an accentual phrase, all "stressed" syllables are necessarily word-final. Thus, the probability of encountering a word-final stressed syllable is always 1, regardless of syllable count in the ISI. There may be exceptions to this pattern in the case of phrase-initial accents (Astésano et al. 2007), but the occurrence of this phenomenon should be randomly distributed with regard to ISI length. French is also a prototypical "syllable-timed" language.

In Finnish, by contrast, stress is always word-initial (Suomi 2007). Thus, monosyllabic ISI in Finnish occur only if a monosyllabic word is followed by any other word (which, in theory, necessarily has initial stress), hence stressed syllables in monosyllabic ISI are

always word-final. In longer ISI, i.e., sequences of a stressed and one or more unstressed syllables, the stressed syllable cannot be word-final, as any unstressed syllables that follows a stressed syllable in Finnish must also fall within the same word. This assumption, too, is certainly overstated – cases of de-stressing, rhythmic beat insertion or weak function words in running speech may result in occurrences of non-word-initial stressed or initial unstressed syllables even in Finnish. Yet, one may assume that Finnish stress assignment rules should lead to a markedly more extreme distribution of word-final stressed syllables with respect to syllable count in the ISI than in English, where words with non-initial or even final stress are not uncommon. Figure 6.7 visualizes the (hypothetical) percentages of stressed syllables that are word-final as a function of syllable count in the ISI for the three languages.

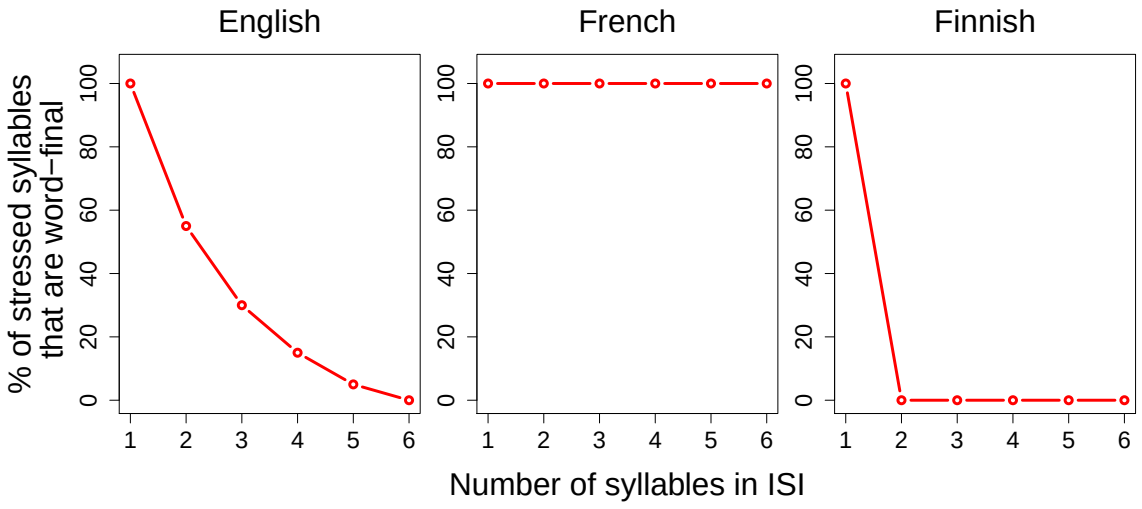


FIGURE 6.7: Percentage of stressed syllables that are word-final as a function of syllable count in the ISI in English (results from MARSEC corpus), French and Finnish (hypothetical distributions; see text for details).

Do these differences in language structure offer an alternative explanation for language-specific differences in "coupling strength ratio" reported by Eriksson (1991) and O'Dell and Nieminen (2001)? In what follows, we will investigate artificial speech data simulated using a minimal "toy model" in order to investigate this question. This toy model will be extremely coarse and ignore many potentially relevant aspects of real speech data. Hence, the goal of our approach will be to supply a proof-of-concept demonstration that, all else being equal, differences in the distribution of word-final (stressed) syllables in a language generate the differences in intercept-to-slope (r) ratios observed in regression analyses of ISI duration by syllable count in various languages, granted that word-final syllables are lengthened. The point of the simulation experiments is not to ultimately

refute O'Dell and Nieminen (1999)'s model, but just to suggest a plausible alternative explanation.

For these simulation experiments, we generated artificial syllable duration data by randomly drawing numbers from log-normal distributions. The log-normal distribution has been found to provide a good approximation of the typically positively skewed distributions of speech segment durations (Rosen 2005) and was therefore considered apt for our purpose. Three categories of syllables were modeled: word-final stressed syllables were drawn from a distribution with a mean $\mu = 265$ ms and a standard deviation $\sigma = 105$ ms, non-word-final stressed syllables from a distribution with $\mu = 178$ ms and $\sigma = 75$ ms, and unstressed syllables from a distribution with $\mu = 147$ ms and $\sigma = 40$ ms. These values were derived from the MARSEC corpus, and, crucially, were used for all three languages. The assumption that stressed and unstressed syllables are of equal duration in English, French and Finnish is of course quite unrealistic, but it was deliberately adopted, as the purpose of the experiment was to isolate the effect of the distribution of word-final lengthening on coupling strength ratios.

The simulated syllable durations were combined into ISI as shown in Table 6.3, in order to match the proportions of word-final and non-final stressed syllables shown in Figure 6.7. For example, the "English" corpus comprised 200 tri-syllabic ISI, 70 of which were created by adding a syllable duration from the word-final stressed distribution to two unstressed durations, while the other 130 tri-syllabic ISI durations were created by adding two unstressed durations to a stressed non-final duration. The simulations thus also incorporate the assumption that ISI of all syllable counts are equally frequent in the three languages, which, too, is not particularly realistic. Many parameters of the simulations could be changed in order to achieve a better representation of the languages under study, but our coarse methodology will suffice for an initial demonstration.

TABLE 6.3: *Assignment of simulated stressed non-final and stressed final syllables by ISI length in English, French and Finnish. The number of unstressed syllables for an ISIS with syllable count n is $200 \times (n-1)$.*

Language	Stress/Position	1-ISI	2-ISI	3-ISI	4-ISI	5-ISI
"English"	Stressed Final	200	110	70	30	10
	Stressed Non-Final	0	90	130	170	190
"Finnish"	Stressed Final	200	0	0	0	0
	Stressed Non-Final	0	200	200	200	200
"French"	Stressed Final	200	200	200	200	200
	Stressed Non-Final	0	0	0	0	0
All languages	Unstressed	0	200	400	600	800

Linear regression models were fitted to ISI duration by syllable count on the simulated data, and the "coupling strength ratio" r, i.e., the ratio between intercept and slope, was computed. This procedure was repeated 500 times for the English and French data. Figure 6.8 shows mean durations and regression coefficients computed over 500 simulations.

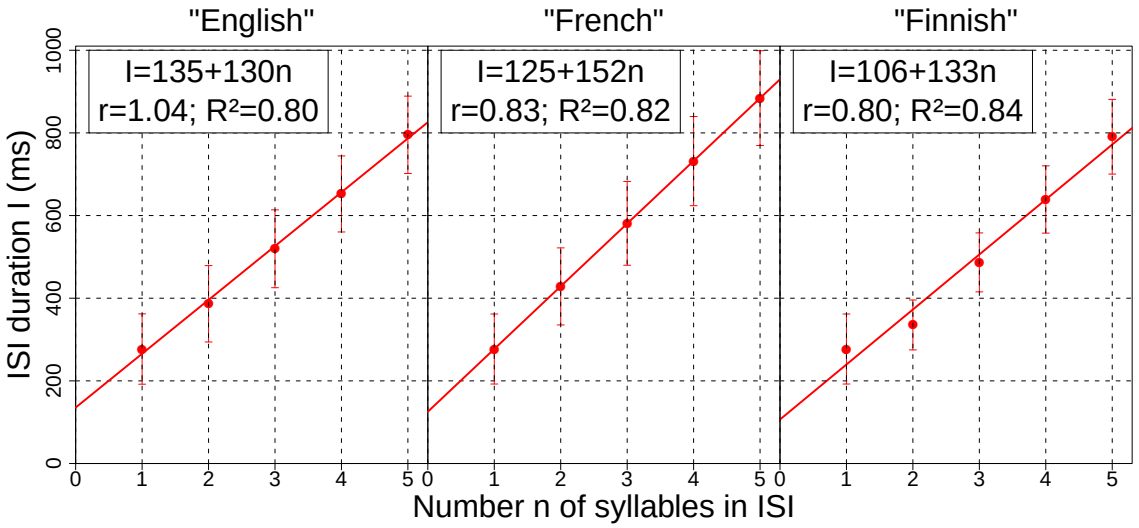


FIGURE 6.8: Regression results on simulated English, French and Finnish data (means, standard deviations and regression coefficients averaged over 500 simulation runs). See text for details.

The simulation result confirms our hypothesis for the comparison between English and the other two languages: all else being equal, the difference in distribution of word-final stressed syllables alone predicts a markedly higher coupling strength ratio for English than for the other languages. For the comparison with French, the explanation lies in the statistical tendency of stressed syllables to be progressively shorter in longer ISI in English (even though ISI length is not the relevant variable as far as timing processes are concerned), which counteracts the increase in ISI duration caused by the addition of unstressed syllables and decreases the regression slope. In Finnish, the abrupt difference between mono- and bisyllabic ISI (100% vs. 0% word-final stressed syllables) makes for a nonlinearity that results in a lower intercept estimate, and, hence, coupling strength ratio compared to English. Interestingly, graphical presentation of Finnish ISI duration means in O'Dell and Nieminen (2001) reveals a very similar nonlinearity. We re-plot the figure from O'Dell and Nieminen (2001) in Figure 6.9 for comparison.

Presentation of O'Dell and Nieminen (2001)'s data in Figure 6.9 indeed suggests that the skewed distribution of stressed syllable durations in Finnish may have something to

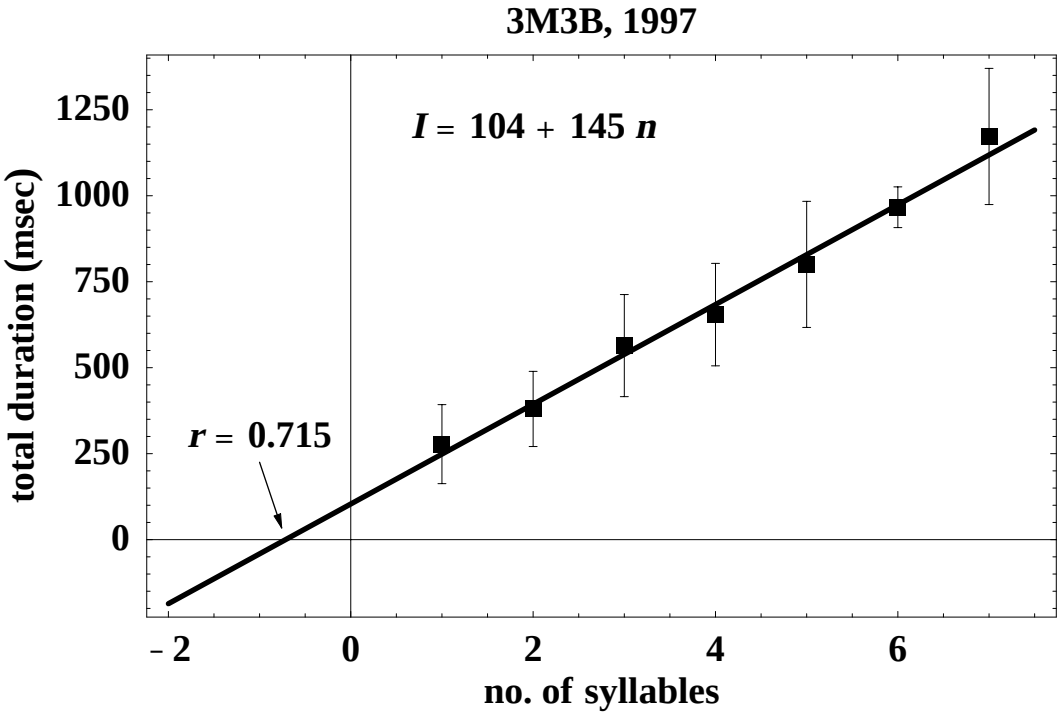


FIGURE 6.9: ISI duration by number of syllables in Finnish (reproduced from O'Dell and Nieminen 2001).

do with the finding of a substantial "coupling strength ratio" despite minimal stressed-unstressed duration differences: note that while O'Dell and Nieminen (2001) report an average stressed syllable duration of 183 ms, the mean duration of stressed syllables from monosyllabic ISI – and, hence, the only word-final stressed syllables in Finnish – is actually as high as 250 ms. Thus, one may hypothesize that there are only few observations from monosyllabic ISI – too few to produce a large mean difference between stressed and unstressed durations, but enough to bias the regression intercept upward. We ran a second simulation study in order to substantiate this hypothesis. In this simulation, we used Finnish-like durational parameters for all three languages, which were chosen so as to match O'Dell and Nieminen (2001)'s durational data more closely: μ(stressed word-final)=265 ms, μ(stressed non-word-final)=175 ms, μ(unstressed)=170 ms. Again, regression analyses were run 500 times.

Results of the second simulation experiment are shown in Figure 6.10. The left panel graphs simulated regression intercepts for the three languages as a function of corresponding mean stressed-unstressed differences. The outcome of the experiment supports our hypothesis: the artificial data with the Finnish language structure yield substantial positive regression intercepts ($>$ 100 ms), while the mean difference between stressed and unstressed syllable duration was only 33 ms on average in the 500 simulations. Admittedly, the simulated stressed-unstressed difference is still markedly higher than

O'Dell and Nieminen (2001)'s 13 ms, but this even more drastic result may be due to other structural factors ignored in our simulation, such as the distribution of pitch accents, or of *unstressed* word-final syllables in Finnish. Whatever the explanation, the tendency predicted by our simulation is correct. This becomes particularly clear in the right panel of Figure 6.10, which graphs ratios between regression intercept and stressed unstressed mean difference computed over the 500 simulations for the three languages: in the Finnish simulations, the average regression intercept is more than three times as large as the difference between mean stressed and unstressed syllable duration. This indicates that the distribution of word-final stressed syllables in Finnish will make for a substantial positive intercept in a regression of ISI duration on syllable count, even if the average stressed-unstressed duration difference is small.

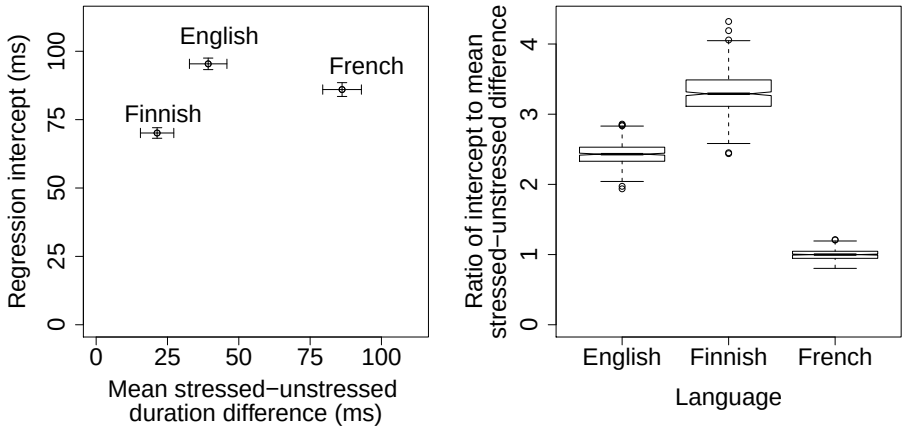


FIGURE 6.10: Left panel: intercepts from regression analyses of ISI duration on the number of component syllables as a function of differences between mean stressed and unstressed syllable durations, computed on simulated English, French and Finnish data (French and English data generated with respective language-specific distributions of word-final syllables, but "Finnish" durational parameters). Right panel: ratios between intercepts and stressed/unstressed differences by language. See text for details.

The ratio between intercept and stressed-unstressed difference is markedly lower in the other languages – in particular, for French, where the proportion of word-final stressed syllables is independent of ISI length, it is exactly one. Thus, the intercept of a regression analysis of ISI duration on syllable count in French should be entirely predictable from the stressed-unstressed duration difference. English, in this analysis, occupies a middle position between French an Finnish, which is not surprising, given that its distributional pattern of word-final stressed syllables is also halfway between the extremes French and Finnish (cf. Figure 6.7). The simulation thus makes two predictions that we can verify on our corpora – the intercept of the regression analysis of ISI duration on syllable count should be markedly higher than the difference between mean stressed and unstressed syllable duration in English, while both quantities should be about the same in French.

This prediction is partially confirmed by our data: analysis of the MARSEC corpus yields a regression intercept of 143 ms for English, while the difference between mean stressed and unstressed syllable duration is only 82 ms. In the French C-PROM data, the regression intercept is 122 ms and the difference between mean stressed and unstressed syllable duration is 94 ms. While the ratio between both numbers is thus not equal to 1 in French, it is at least substantially lower than in English (French: 1.30, English: 1.74).

6.3.3 Discussion

These results show that differences in the relative frequency and distribution of stressed syllables that are subject to word-final lengthening in different languages may contribute to observed differences in the way ISI duration depends on the number of component syllables. Such differences thus provide an alternative to coupled oscillators in accounting for results by Eriksson (1991). Our very coarse simulations did obviously not provide perfect matches to empirical results, and there may be many other structural factors not implemented in our toy model that influence the relationship between ISI duration and the number of component syllables. However, our analysis has created a *ceteris paribus* condition, showing that all else being equal, the structural differences we modeled will trigger tendencies in the observed direction.

The patterns uncovered in our analysis of course do not necessarily account for the lower coupling strength ratios for languages such as Italian, Spanish and Greek compared to English reported by Eriksson (1991). We are not aware of detailed information on the distribution of stressed syllables with regard to word boundaries in these languages. As for Italian, Krämer (2009) reports that only about 2% of all words in Italian have final stress, which would suggest that the distribution of word-final stressed syllables in Italian follows a Finnish-like pattern. This figure refers to word *type* counts, though, and it is not clear how frequent word forms with final stress are in spoken Italian in absolute terms. In Spanish, final stress does seem to be quite frequent (Eddington 2000). In any case, the lower coupling strength ratios reported for these languages compared to English may stem largely from lower stressed/unstressed syllable duration ratios in these languages, due to the absence of factors such vowel reduction in unstressed syllables or quantity-sensitive stress assignment. Moreover, word-final lengthening of stressed syllables may be weaker or absent in these languages – for example, recall that d'Imperio and Rosenthall (1999) report that at least in open syllables, stressed vowels are actually *shorter* word-finally than in other positions in Italian. As for the languages considered in our study, word-final lengthening (or at least a lengthening effect in a constituent below the intonational phrase level) is supported for English by our own results and for Finnish by results reported in Suomi (2007). For French, the question of

word-final lengthening in stressed syllables is moot, as there are no stressed syllables in other positions.

Whatever the explanation, our simulation study has shown that structural factors may play a key role in explaining Eriksson (1991)'s result. While oscillatory models account for the data in a highly elegant way, we have shown that the result may be explained by an arguably simpler alternative.

6.4 General Discussion

In this chapter, we have investigated two predictions of models of speech timing concerning effects of the syllable count in the ISI on syllable and ISI duration. Our optimization-based model of speech timing makes no specific predictions regarding both phenomena, and our analyses suggest that it is justified to do so: the observed statistical tendency of syllables to shorten as a function of the syllable count in larger prosodic units is an artifact of the decreases probability of observing word-final and hence lengthened syllables in units with greater syllable count. This effect also provides a possible explanation for results on ISI expressed as a function of the number of component syllables. We have seen that our model will reproduce the statistical patterns simply by incorporating word-final lengthening. As we said above, the explanation in this case does of course not lie in the model architecture, but in distributional patterns of languages.

Our results are interesting with regard to current debates about speech timing mechanisms: they support the domain-and-locus approach to speech timing (White 2002, 2014), which argues that suprasegmental speech timing is confined to localized lengthening effects at the heads and edges of prosodic domains, and does not include quasi-periodic compensatory mechanisms, as implied by some of the models reviewed in Chapter 4. This is particularly true in view of the contrastive rhythm effects we observed in the analysis of the MARSEC corpus. Our model does currently not provide a unified explanation of these effects. A model that is capable of integrating all these phenomena would be a desirable achievement for further research. f

Chapter 7

Incompressibility

7.1 Introduction

We have encountered the hypothetical speech timing property of incompressibility several times in this work. In this chapter, we will first report an empirical study that attests this phenomenon as a property of suprasegmental constituent durations in speech, and then present simulation experiments demonstrating that the phenomenon automatically emerges from the architecture of our model. Incompressibility receives separate treatment because we classify it as a more low-level property of speech timing related to, in contrast to high-level linguistic effects on speech timing, which will be treated in the subsequent chapter. In fact, our model will be shown to implement the hypothesis that some of these effects are based on incompressibility.

As discussed above, the term incompressibility refers to the intuition that speech sounds cannot be arbitrarily short, i.e., that there are lower limits to their durations, for articulatory and perceptual reasons. The concept was made popular by Klatt (1973), who studied the combined effects of polysyllabic shortening and (absence of) postvocalic voicing on vowel duration in presumably accented words. He found that both factors in combination shorten a target vowel in an otherwise constant environment less strongly in both absolute and proportional terms than would be predicted by simply adding the effects of both factors in isolation: the difference in /ɛ/ duration between *bed* and *betting*, for example, is smaller than would be predicted from adding the difference in /ɛ/ duration between *bet* and *bed* to that between *bed* and *bedding*. Klatt conjectured that this is due to incompressibility – shortening effects diminish as the hypothesized lower duration threshold is approached – and captured this in the following descriptive model of vowel duration:

$$D_j = k(D_i - D_{min}) + D_{min} \tag{7.1}$$

The model starts from some *inherent* duration D_i that is equal to a vowel's duration in an isolated, pitch-accented monosyllable, i.e., something like its longest conceivable instance, and then proceeds by recursively applying shortening factors $k < 1$ for individual timing processes, resulting in the output duration D_j. For a given factor, D_j is used as input duration D_i in the next iteration. Incompressibility is invoked by adding the constant D_{min}, and by applying k only to the term $D_i - D_{min}$, i.e., to the compressible part of vowel duration. This accounts for the less than additive shortening effect of several factors in combination: the multiplicative shortening factor is applied to a progressively shorter base duration.

Klatt's model can be used to predict vowel duration, given that the free parameters are appropriately estimated from data. At the same time, it provides a simple analysis method for attesting incompressibility: Equation 7.1 has a straightforward linear regression form, as is immediately apparent from Figure 7.1. Thus, incompressibility as a property of a given timing factor can be investigated by regressing durations of vowels characterized by different levels of that factor (such as voiced vs. voiceless postvocalic consonant or monosyllabic vs. bisyllabic word) on each other and testing whether the regression intercept is statistically different from zero. This methodology was applied in an empirical study, which we will now report.

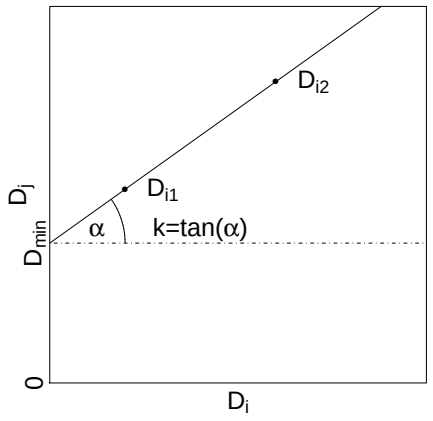


FIGURE 7.1: Graphical illustation of Klatt (1973)'s duration model. Input duration D_i (exemplified by two hypothetical concrete durations D_{i1} and D_{i2}) is mapped onto output duration D_j by multiplying the compressible part $D_i - D_{min}$ by some factor k and adding the incompressible constant D_{min}.

7.2 Corpus analyses

7.2.1 Data and Method

In this chapter, we will investigate incompressibility as a property of *syllable* durations under speaking rate variation, using data from two speech corpora with somewhat complementary characteristics. This will be done using the methodology introduced above, that is, by regressing syllable durations from fast-rate productions on corresponding syllable durations from slower rate productions, and taking the regression intercept as an estimate of incompressibility. Incompressibility as a property of speaking rate effects on *segment* duration has been established in combination with other effects, such as vowel tensity and postvocalic voicing (Gopal and Syrdal 1987, Port 1981). Our analysis, though conducted on syllable durations, will thus not provide dramatically new fundamental insights, and the regression method sketched above has also been applied previously in similar ways (e.g. Cummins 1999, Gopal 1996). The point of the analysis is simply to give a further demonstration of the effect and to corroborate previous results, and, ultimately, to provide data that our syllable-centered model is able to simulate.

The first of the two analyses was conducted on data from two corpora, the *BonnTempo Corpus* (BTC; Dellwo et al. 2004) and a single-speaker database compiled for corpus-based speech synthesis, which will be described in more detail below. The BTC comprises readings of a short paragraph of text obtained from different numbers of speakers of German, English, French, and Italian in their respective native language. Dellwo et al. (2004) report that categorical variation in overall speaking rate was induced in the production of the corpus by first prompting speakers to read the text at a "normal", spontaneously adopted speaking rate, and then to repeat it four times, at two degrees of acceleration and deceleration relative to the "normal" condition, respectively. This yields five tempo conditions: *very slow, slow, normal, fast* and *fastest possible*. We restricted our analysis to the latter three conditions, regressing syllable durations from the fast rate conditions on the corresponding durations from the normal rate condition. Table 7.1 summarizes the corpus data.

Separate regression models were fitted to stressed and unstressed syllable durations, using the existing annotation and prosodic transcription of the corpus. Phrase-final syllables were excluded from the analysis, in order to prevent a possible confounding influence of final lengthening. As can be seen from Table 7.1, syllables were sometimes elided, i.e., not realized at all in the fast conditions. These syllables were also excluded

TABLE 7.1: *Syllables and speakers per language and actual speaking rate by tempo condition in the BonnTempo Corpus. Syllable counts are token frequencies pooled across speakers. Numbers in brackets denote numbers of elided syllables in the fast/fastest possible condition.*

Language	Speakers	Stressed syllables	Unstressed syllables	Speaking rate (syls/s)		
				normal	fast	fastest
German	15	408 (0/1)	609 (0/0)	5.4	6.1	8.8
English	7	182 (0/0)	304 (4/14)	5.8	6.4	8.0
French	6	134 (0/0)	311 (0/1)	6.1	6.8	9.3
Italian	3	107 (1/1)	180 (4/18)	7.1	7.9	11.1

from statistical analysis,[1] but they will be retained in the plots as cases of zero duration in the fast conditions.

7.2.2 Results

Table 7.2 summarizes the regression models, showing estimates (in ms) and significance levels for intercepts (Int) as well as the amount of variance (R^2) explained by the models (**: $p < .01$; ***: $p < .001$; all slopes are significant at $p < 0.001$). Figures 7.2 and 7.3 show plots of syllable durations at the fast rates as a function of durations at normal rate in the four languages, with regression lines fitted to the data.

TABLE 7.2: *Summary of regression models of stressed and unstressed syllable duration from fast/fastest possible productions on corresponding durations from "normal" rate productions in German, English, French, and Italian data from the BonnTempo Corpus.*

	Fast		Fastest possible	
Language	Int	R^2	Int	R^2
Ge str.	39***	0.71	40***	0.47
Ge unstr.	17***	0.72	16***	0.57
En str.	21**	0.77	43***	0.55
En unstr.	16***	0.66	19***	0.49
Fr str.	32**	0.52	54***	0.31
Fr unstr.	18***	0.71	13**	0.54
It str.	26**	0.68	31***	0.60
It unstr.	30***	0.58	23***	0.42

The overall pattern of results of the analysis is clear-cut: significantly positive intercept estimates are consistently observed throughout the corpus, regardless of language and rate condition. The data thus provide substantial evidence for the hypothesis that increasing overall speaking rate measured in terms of syllable duration is characterized by

[1] In Windmann et al. (2013), elided syllables were included in the analysis, which leads to slightly different results in some cases, the intercept estimate in the fastest-possible unstressed Italian data turning out non-significant. We now feel that it is more appropriate to exclude elided syllables, as they constitute massive outliers. In any case, the general pattern of results is the same.

incompressibility. Again, we argue that no correction for multiple analyses is necessary given the highly consistent pattern of results. One may observe that there is a tendency for the models fitted to stressed syllables to have higher intercepts than their unstressed counterparts. This tentatively supports the assumption that unstressed segments – or, in this case, syllables – are more compressible than stressed ones (Klatt 1979). Since the well-established correlation between stress and duration made it impossible to assess this claim statistically in combined models with stress as a predictor, this observation has to remain impressionistic.

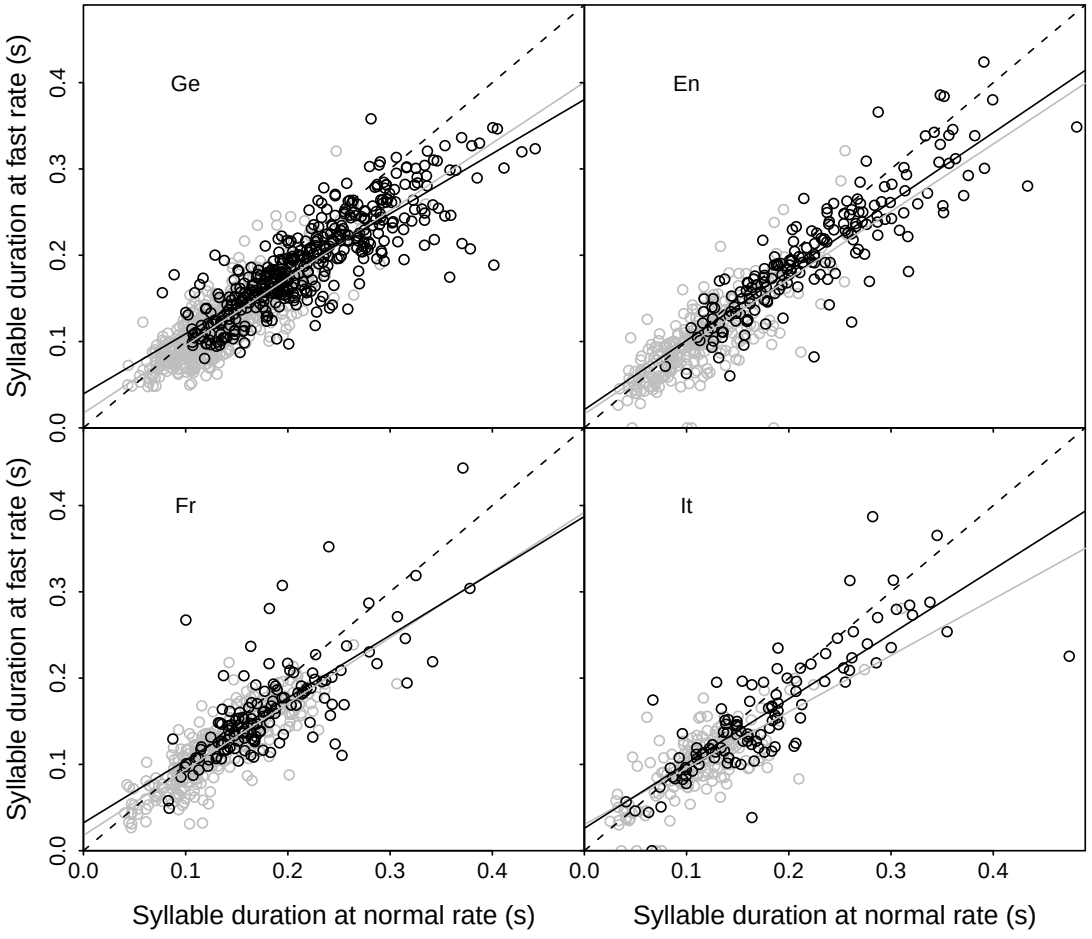


FIGURE 7.2: Stressed (black) and unstressed (gray) syllable durations from fast productions plotted as a function of corresponding syllable durations from "normal" rate productions in the BonnTempo Corpus, with fitted regression lines, in German, English, French and Italian.

The data from the BonnTempo Corpus provide evidence for incompressibility of syllable durations as a function of speaking rate variation across various languages, stress conditions and for different levels of rate variation. However, the corpus is quite limited in some other respects. One obvious limitation is its small size. The syllable counts

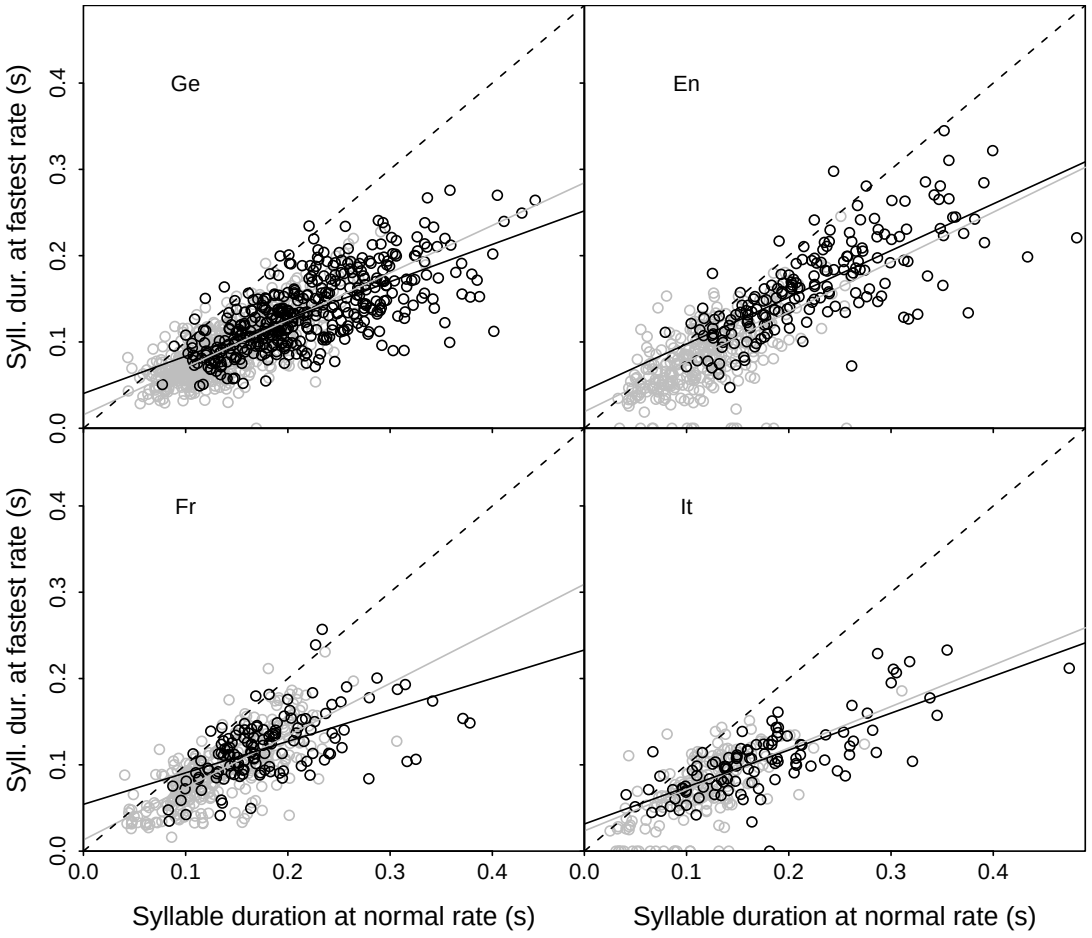


FIGURE 7.3: Stressed (black) and unstressed (gray) syllable durations from fastest possible productions plotted as a function of corresponding syllable durations from "normal" rate productions in the BonnTempo Corpus, with fitted regression lines, in German, English, French and Italian.

in Table 7.1 are token counts pooled across speakers. The syllable *type* count is between 70 and 100, depending on the language. Moreover, materials are not controlled in any way, and there is also between-speaker variation. While these characteristics constitute potential risk factors for positing spurious effects, we would argue that this is unlikely in the present case, given the very consistent pattern of results observed across different partitions of the corpus data. Nevertheless, we will supplement our argument with the analysis of data from a second corpus, which to a certain extent addresses the shortcomings of the BTC analysis.

This additional analysis was carried out on a speech database that we will refer to as the *Petra Corpus* (PC). This corpus consists of 400 utterances produced by a single trained female speaker of German at a slow (average rate of 4 syls/s) and a fast (average rate of 8 syls/s) speaking rate for the purpose of integrating fast speech in a unit selection speech synthesis system (Moers et al. 2010). This resource is thus obviously limited to

one speaker and one language, and, moreover, to a rather special speaking style: the speaker had to articulate as clearly as possible in the fast rate condition in order to produce fast yet intelligible speech for speech synthesis integration. The fast PC data could thus arguably be classified as a form of "fast hyperspeech". However, compared to the BTC, the PC has the benefits of eliminating between-speaker variation, and of containing enough data to facilitate controlling for the segmental identity of syllables, in addition to stress level and within-utterance position.

We determined the frequencies of the individual syllable types in the corpus, "syllable type" here referring to any given set of non-utterance final syllables sharing the same sequence of phonetic segments and stress level. For this analysis, we implemented an outlier removal procedure, excluding syllables whose duration exceeded the mean duration in either rate condition by more than 2.5 standard deviations, as outliers might influence results considerably given the smaller cell sizes in this analysis. Only syllable types with at least 30 remaining tokens after these data selection procedures were retained in the study. 19 syllable types from the corpus, mostly highly frequent function words and grammatical affixes, were found to match this criterion. The *lmList()* function from the *lme4* package (Pinheiro and Bates 2000) in *R* (R Core Team 2014) was used for regressing fast-rate on corresponding slow-rate syllable durations separately within individual syllable types in these data. The *lmList()* function allows for specifying a grouping factor, and fits separate regression models within the levels of that factor. *lmList()* estimates residuals based on pooling across the whole dataset and therefore has greater statistical power than fitting separate regression models manually using R's *lm()* syntax. The general strategy of fitting separate models within syllable types was chosen because comparisons among the different syllable types are obviously not meaningful. Results are listed in Table 7.3 and presented graphically in Figure 7.4.

In contrast to the BTC, no elisions were observed in the Petra Corpus, which is of course a consequence of the rather special speaking style. The statistical analysis of the PC mirrors the results from the BTC analysis: significantly positive intercepts are found for all syllable types, indicating incompressibility as a function of rate variation. In these data, however, slopes failed to reach significance in a number of cases. Different explanations for this outcome could be conceived. One possibility is that the speaker articulated these syllables so rapidly that all durations were indeed reduced to the compressibility threshold. Alternatively, the failure to observe significant slopes in some cases may simply be due to a lack of statistical power. If this were true, however, one would expect that non-significant slopes are mostly observed for the least frequent syllable types, which is generally not the case. Whatever the explanation, the important result is that significant intercepts are observed throughout.

TABLE 7.3: *Regression summary of fast rate on slow rate durations by syllable type in the Petra Corpus (.: $p < 0.1$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.0001$).*

Syllable	Intercept (ms)	Slope	Number of tokens
/ʔaɪ/	37*	0.32***	56
/ʔaɪm/	53*	0.32***	45
/bə/	48***	0.25*	55
/də/	35***	0.37***	40
/dɐ/	76***	0.07 n.s.	46
/daːs/	93***	0.18*	54
/deːɐ/	42**	0.28**	90
/deːn/	73**	0.21*	34
/diː/	60***	0.17**	119
/fɛɐ/	51*	0.30*	42
/gə/	36**	0.37**	86
/ʔɪm/	76***	0.15*	79
/mɪt/	100***	0.06 n.s.	32
/nə/	58***	0.22 .	44
/nən/	87**	0.09 n.s.	39
/ʁə/	34*	0.34*	32
/tə/	73*	0.09 n.s.	88
/tən/	90***	0.02 n.s.	60
/ʔʊnt/	126***	0.00 n.s.	81

7.2.3 Discussion

Our two corpus analyses support the assumption that the effect of changes in overall speaking rate (of *increases* in overall speaking rate, to be specific) on syllable durations exhibits incompressibility: the positive regression intercepts provide an empirical estimate of Klatt's D_{min}, which may of course vary due to additional factors such as the segmental makeup of the syllable or its stress level. Our results thus complement earlier findings made at the segmental level. A possible objection against our analysis is that incompressibility is in fact not a property of syllables, but of their component segments. This may be true, but with a view to the architecture of our model, we would argue that it is not a crucial point: as we said above, the choice of syllables as the level of representation in our model is not meant to raise strong claims about syllables as units of speech production or cognitive representation; they are rather viewed as a proxy for the included segments or gestures, assuming that the efficiency principles the model is based on should apply at lower linguistic levels in a similar fashion.

In the following section, we will report on simulation experiments that demonstrate the replication of the basic incompressibility pattern in our model. In the arguably more naturalistic BTC data, we have observed that in fast speech, entire syllables may be

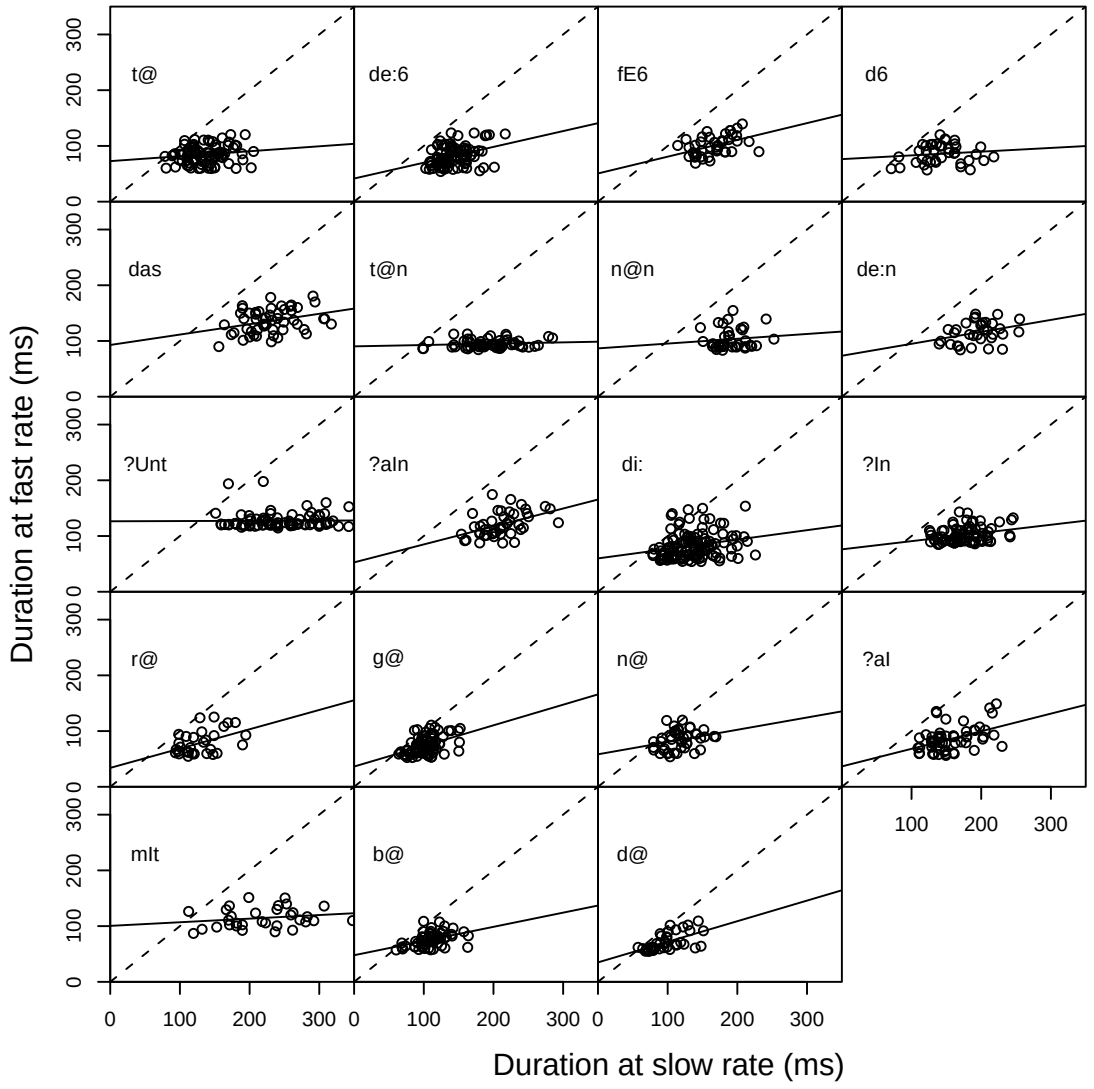


FIGURE 7.4: Fast∼normal regression analyses on syllable durations within individual syllable types (transcriptions coded in SAMPA) in the Petra Corpus.

deleted in the acoustic domain. As we said above, we deliberately chose a model architecture that does not allow deleting syllables, as we do not feel comfortable engaging in the discussion on the reality of deletions, which requires consideration articulatory information. We will supply a somewhat more speculative account of the deletion phenomena observed in the BTC later in this chapter, using a modified model that does allow for deleting entire syllables.

7.3 Modeling Durational Incompressibility

7.3.1 Basic Model

In order to demonstrate the replication of the effect in our model, we compiled a "corpus" of 40 sequences of ten syllables each (five stressed and five unstressed), which was used in two simulations, with different settings of the α_D parameter simulating increasing global speaking rate. There are several reasons why we did not attempt at matching more closely the characteristics of any of the corpora used in the empirical study. First, as stated in the Introduction to this work, the purpose of our model is not to approximate any given set of real-world data as closely as possible, but to derive principled explanations for empirically observed phenomena. The basic facts about incompressibility have been observed in all of the datasets in the corpus study, so that a generic "toy corpus" should suffice to demonstrate the general effect. Moreover, the individual datasets from the corpus study, e.g. the language-specific sub-corpora of the BTC, are not directly comparable to each other, and one cannot know whether observed differences are a function of language-specific implementation of the timing processes involved, or if they simply reflect structural properties of the particular datasets. We also have no principled a priori hypotheses about possible language-specific parameter settings.

For the initial simulation experiment, we ran the model twice, with $\alpha_D = 1$ and 2, respectively, to simulate a slower and a faster production of the corpus. Some durational variation was introduced by drawing the value of the η_i parameter from a normal distribution with a mean of 1 and a standard deviation of 0.2 for each syllable. The accentual lengthening parameter Ψ_W was fixed at 0 in these simulations, in order to keep the model setup simple for our basic demonstration. The other parameter settings were the same as those used for the simulation depicted in Figure and described there.

Results of the simulation are shown in Figure 7.5. The first thing to notice is that the simulation does not provide a particularly close match of the empirical data; for example, there is no overlap in stressed and unstressed duration distributions, and the regression lines also do not cross, as seems to be the case with most of the empirical data shown in Figures 7.2 and 7.3. Both outcomes are consequences of the particular noise parameters used in the simulation, however, and are not central to the issue at hand. What is crucial is that the regression analyses on the model output confirm that the model successfully reproduces the basic pattern of results found in the empirical data: regression intercepts are significantly greater than zero for both stressed ($t = 17.31, p < 0.0001$) and unstressed syllables ($t = 13.43, p < 0.0001$). The stressed intercept is also greater than the unstressed one, in accordance with most of the empirical data summarized in Table 7.2.

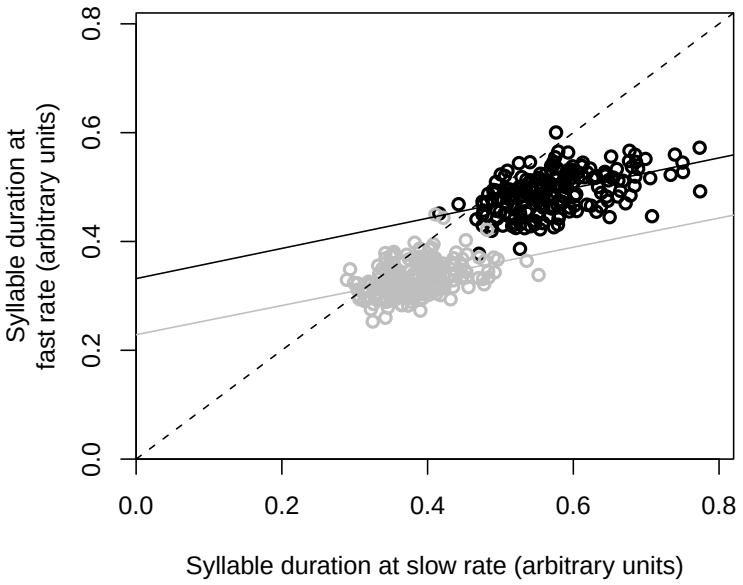
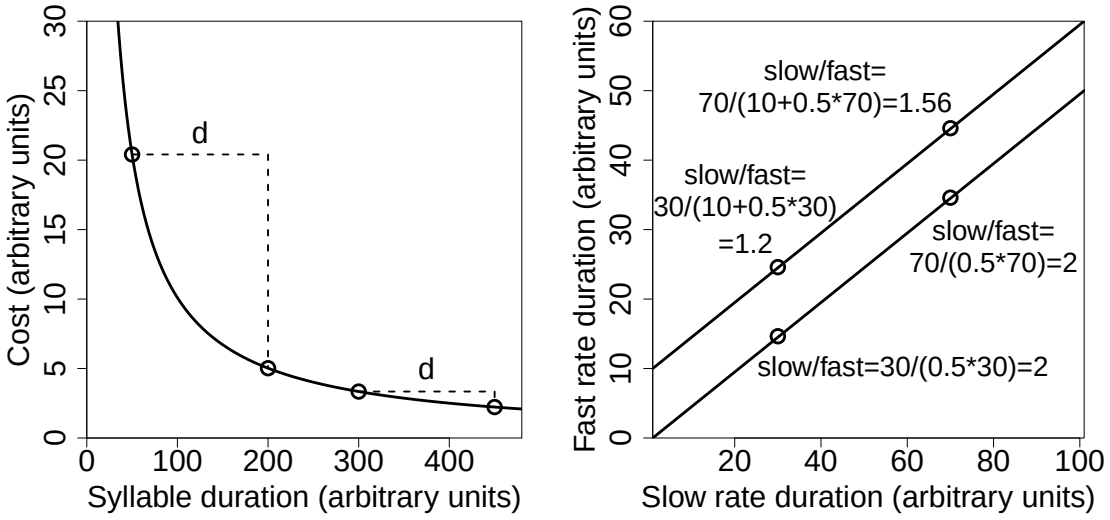


FIGURE 7.5: Simulated syllables durations at fast ($\alpha_D = 2$) as a function of simulated syllable durations at slow ($\alpha_D = 1$) rate. See text for details.

The replication of the incompressibility pattern is a consequence of the mathematical properties of the perceptual cost function P_s, which is illustrated in panel (a) of Figure 7.6: shortening progressively shorter syllables by a constant amount of duration will result in progressively higher cost. The connection to the positive regression intercepts becomes apparent once we recall the original motivation of the concept of incompressibility as formulated by Klatt (1973), i.e., to account for the fact that some speech timing factors in combination have a smaller influence on duration than would be predicted from adding their durational effects when applied in isolation. Indeed, a different way to describe the pattern in Figure 7.5 (or in the real data in Figures 7.2 and 7.3, for that matter) is to say that longer syllables shorten proportionally more strongly in fast speech than shorter syllables. This is shown in panel (b) of Figure 7.6: due to the positive constant in the regression equation, the slow/fast duration ratio increases with duration at the slow rate. By contrast, with a zero intercept, i.e., without incompressibility, the slow/fast ratio stays constant, regardless of slow rate duration.

Incompressibility being a consequence of the perceptual component of our model, the explanation of the effect suggested by it would be that extremely short durations are perceptually too costly. The replication of the effect is a quite obvious consequence of the model architecture, and a critical observer might ask whether the P_s function does not pretty much "hardcode" the effect into the model, in the same way Saltzman et al. (2008) encode the pattern observed in the data of Kim and Cole (2005). The crucial difference, however, is that in our model, the design of the responsible component P_s is based on independently motivated principles, as detailed in Chapter 5. Saltzman et al.



(a) (b)

FIGURE 7.6: Illustrations of the incompressibility effect in our model. Panel (a): the same durational reduction d incurs a higher increase in perceptual cost for a short than for a longer base duration. Panel (b): hypothetical slow-fast regressions, showing that incompressibility (i.e., a positive intercept) implies increasing slow/fast ratio for increasing slow duration, whereas non-incompressibility (zero intercept) implies constant duration ratio between slow and fast condition.

(2008)'s coupling strength modulation lacks such independent motivation; the authors do not even seem to be particularly concerned about this fact. In our opinion, a model that purports to be explanatory should aim for higher goals than merely demonstrating that various things can be *implemented* in it.

This being said, it is clear that our model only represents a first-pass approximation of the empirical data. For one thing, the particular value of the regression intercept observed in our above simulation does not represent a hard lower boundary, but depends on the values of the model parameters. Experimentation with increasing values of the speaking rate parameter α_D showed that for extreme settings of this parameter, durations at the fast rate get arbitrarily close to zero, but the regression intercept still stays significantly positive. Regression slopes, by contrast, quite quickly approach zero when α_D is increased, replicating the pattern found in the Petra Corpus and suggesting that cases where non-significant slopes were observed in these data indeed mark cases where the speaker reached some physical boundary in the fast condition, articulating everything as quickly as possible. As for the behavior of the model at extreme α_D settings, we simply have to stipulate that such settings are meaningless, as speakers cannot talk infinitely fast due to physical limitations of their vocal tracts, which are not explicitly accounted for in our model.

7.3.2 Modified Model

One further observation that our basic model has nothing to say about are the occasional cases where syllables were entirely deleted in the acoustic domain in the BTC data. As discussed above, deletions in speech are a controversial issue, and a truly satisfactory treatment may require fine-grained articulatory analysis. Nevertheless, we have developed a modified version of our model that does allow for deleting syllables. With this version of the model, we assume that no matter how apparent deletions in the acoustic domain are implemented at the gestural level, they may be a consequence of higher-level timing constraints. We will see that this version of the model achieves quite interesting results, even though we will see that it makes an incorrect prediction regarding prosodic prominence.

We re-plot the relevant data from the BonnTempo Corpus, zooming in on unstressed syllable durations in the fastest-possible condition in English in Figure 7.7 below in order to more clearly demonstrate the effect. It is apparent that deletion of entire syllables in the acoustic domain is quite frequent, which would seem to argue against the notion of incompressibility as a lower duration boundary. However, one may observe that in a statistical sense, incompressibility is still valid: note that while deletions do occur, durations in the fast condition do not smoothly approach zero; they seem to literally "fall from the ceiling" at the regression intercept. Accepting for a moment the premise that "true" deletion of syllables exists, we may propose the following hypothesis: incompressibility is not a hard lower boundary; rather, speakers have two options: either to produce a syllable with at least the minimum duration D_{min} (empirically estimated by the regression intercept), or to delete it. Syllables with positive durations $< D_{min}$ would lie in the incompressible region and therefore cannot be produced.

It is clear that for the model to reproduce this pattern, a perceptual cost function with a finite intercept is needed, so that zero durations are not punished by infinite cost. In order to achieve this, we redefined P_S as follows:

$$P_S = \sum_i e^{-\psi_i s_i}. \tag{7.2}$$

The choice of this function over Šimko (2009)'s one plotted in Figure 5.4 is arbitrary, and experimentation revealed that both functions make qualitatively similar predictions. Figure 7.8 plots the function for two values of the stress parameter ψ_i. With the perceptual cost redefined as in equation 7.2, we have to use *smaller* absolute values for ψ_i to increase the function value of P_S and thus the prominence of a syllable, as the exponent in equation 7.2 is negative.

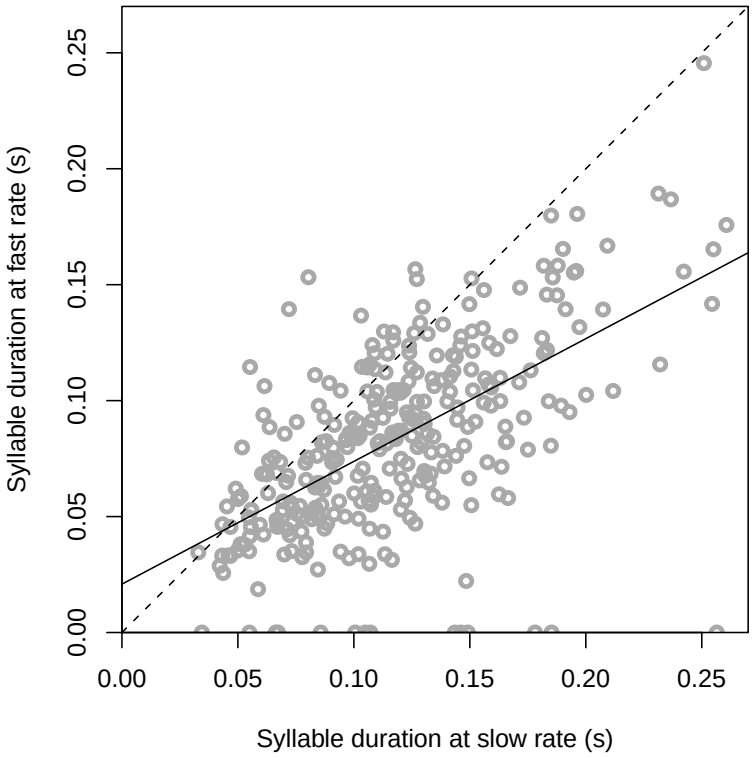


FIGURE 7.7: Fastest possible∼normal regression on English unstressed BTC data.

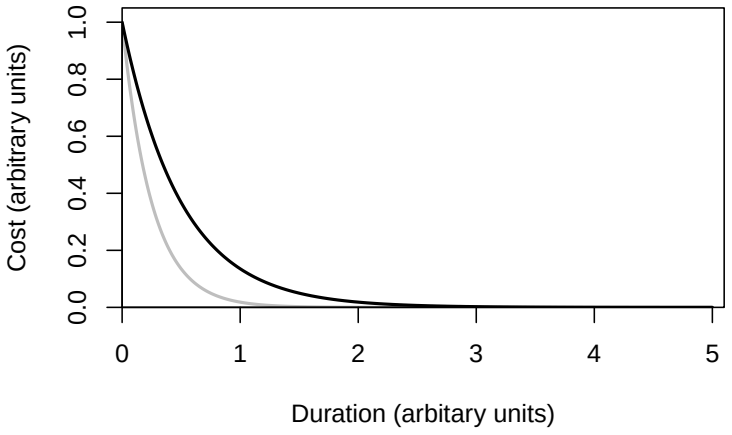


FIGURE 7.8: Modified perceptual cost function $P_S = e^{-4s_i}$ ("unstressed", gray) and $P_S = e^{-2s_i}$ ("stressed", black).

The function in equation 7.2 has a finite intercept – it is equal to 1 – and will thus presumably allow the model to delete syllables. Note, however, that the replication of incompressibility as defined above is far from trivial: the challenging aspect of the data is not the possibility of deletions as such, but rather the fact that D_{min} represents a kind of discrete boundary, with durations smoothly shortening up to this point, but instantaneously jumping down to zero if even stronger impetus towards shortening is applied. We may borrow a term from dynamical systems theory here and classify this pattern as a *bifurcation*, a point where a small continuous change in some parameter

leads to a sudden qualitative change in system behavior (cf. Šimko et al. 2014b).

We repeated the simulation with the modified model, using somewhat different parameter settings. We set $\alpha_E = 3$ and $\alpha_P = 5$. η_i was drawn from normal distributions as above. $\delta_i = 1$ was used for all syllables. $\alpha_D = 1$ was used in the "slow" $\alpha_D = 6$ in the fast simulation run. The accent parameter Ψ_j was set to zero. Results of the simulation experiment are shown in Figure 7.9. No lexical stress distinction was introduced here, setting ψ_i to 4 for all syllables.

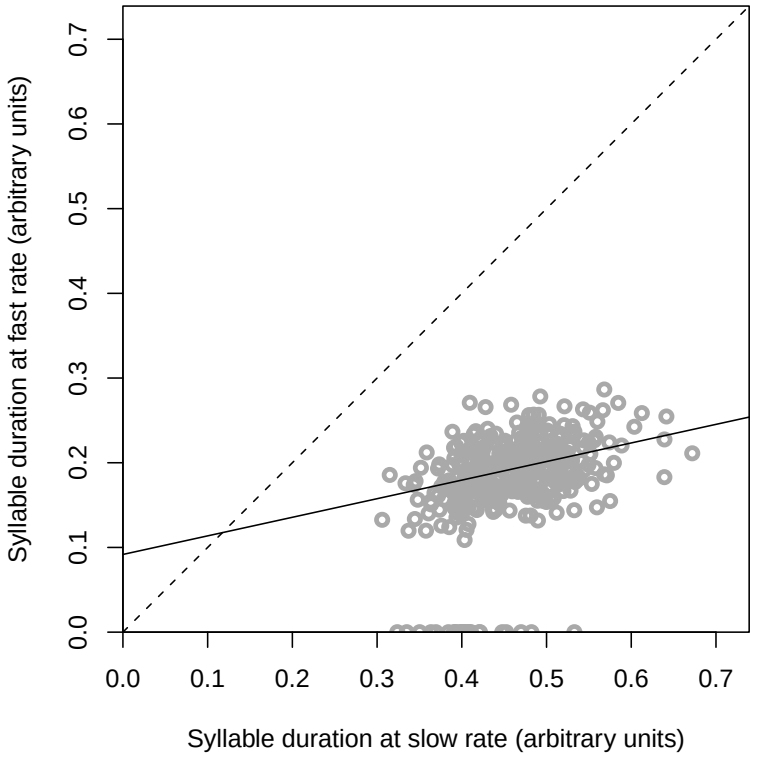


FIGURE 7.9: Simulated syllables durations from the modified model at fast ($\alpha_D = 6$) as a function of simulated syllable durations at slow ($\alpha_D = 1$) rate. See text for details.

We may note that again, the numerical match between real and simulated data is not particularly close. We could have tried to find parameters so as to achieve a closer fit to the data, but this was not our intention. What is important is that the simulation reproduces the key qualitative property of the natural data: the emergence of the incompressibility bifurcation, as is evident from the existence of deletions despite the positive regression intercept ($p < 0.001$, computed on non-deleted durations), and the absence of durations that lie in the "incompressible region" between zero and the regression intercept.

The model's simplicity allows us to demonstrate how this behavior arises. For the given settings, optimal solutions can in fact be found by simply graphing cost function C for an individual syllable and visually identifying the minimum. This is possible

because given that the word-level perception cost (accent parameter Ψ_j) is set to zero and overall duration cost D is simply linear, the optimal solution for any given syllable is determined completely locally, and does not depend on other parts of the simulated utterance.[2] Figure 7.10 plots the cost function C for a hypothetical syllable, with the parameter settings used in the "fast" simulation reported above, but ψ_i fixed at 4 and η_i (the local weight for component cost function E) increasing from 1 to 1.75 in increments of 0.05. The individual trajectories correspond to different values of η_i. The circles in the plot mark the minima of C for the individual η_i values, corresponding to optimal durations. The dashed line marks the intercept of C. As expected, the optimal duration initially smoothly decreases with increasing η_i (as more and more premium is placed on speaker's effort, leading to progressively "lazier" production). At some cutoff point, however, the trough in the cost function C crosses its intercept, and optimal duration instantaneously jumps down to zero. The crucial feature of the model that is responsible for mimicking the empirically observed deletion pattern is the emergence of the little "hump" close to the intercept of the cost function, due to which the cutoff point lies at a positive duration value.

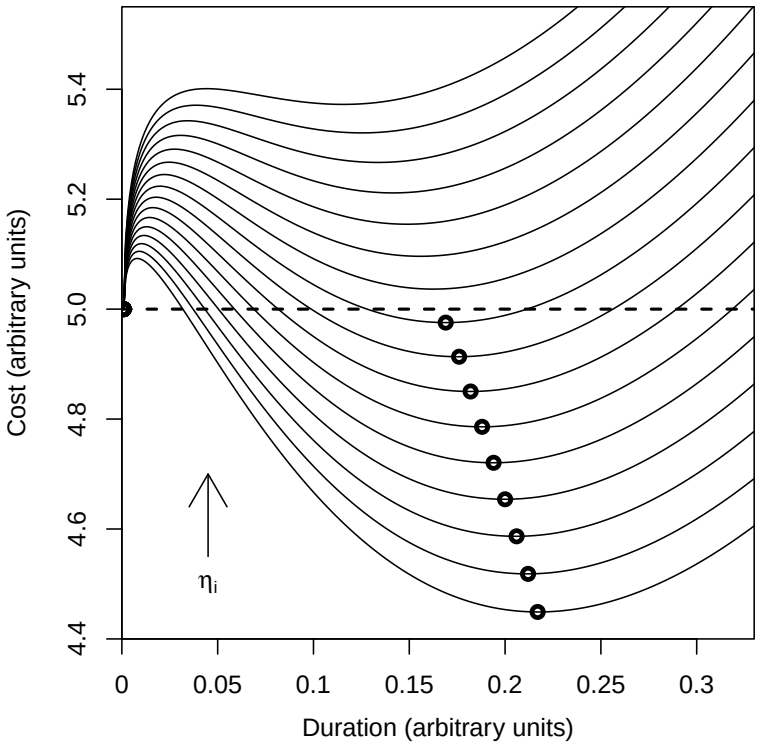


FIGURE 7.10: Plot of modified cost function C for different values of η_i, with parameter settings as given in Section 3.1. Black circles mark optimal durations for a given η_i value, and the dashed line marks the intercept of C. See text for more details.

[2]We can safely assume that cost function C has no other minima outside the plotting range, since D and E are monotonically rising, and the only falling component, P, eventually levels out.

Thus, incompressibility emerges from the interaction of the different component cost functions in our model, without any explicit ad-hoc mechanisms. It is the interplay of the three components of C that is responsible for the replication of the effect. For very short durations, the steep rise of the overall cost function C reflects the initial rapid rise of E. As duration increases, however, the influence of E on overall cost dwindles, as its slope eventually decreases, and the perceptual cost function P_S begins to dominate the evolution of the overall cost function C. This continues until the minimum is reached. Beyond the minimum, overall cost starts to rise again, as P_S approaches zero.

This shows how the interplay of the individual component cost functions triggers the emergence of an incompressible duration region. The independent assumptions we used to motivate the individual component cost functions suggest a straightforward interpretation of the effect: very short durations would require that the speaker produce extremely fast articulations (within certain physical limits) or massive undershoot of articulatory targets, leading to degradation of the acoustic cues to the identity of the intended propositional content. In either case, the gain on P_S achieved by producing a very short (as opposed to zero) duration is not great enough to offset the costs on the E and D dimensions incurred by hypothetical durations in the incompressible region. As producing such durations would entail spending effort while no perceptual benefit is generated, the better option in this case is to remain silent and not to produce the syllable at all, so that at least effort can be saved while the fulfillment of perceptual requirements (for this particular syllable) is impossible. The concave shape of E is obviously crucial for the effect to be borne out. We regard this dependence as unproblematic, as the usage of a concave function (the square root of syllable duration) for E is not arbitrary, but based on independently motivated assumptions, as detailed in Chapter 5.

We investigated the model behavior under increasing α_D values in the fast condition. It turned out that with extreme settings of the rate parameter, solutions of the optimization problem become essentially unstable – running the modified model several times with identical parameter settings ($\alpha_D = 11$) produced completely different results, even though the optimization procedure did seem to converge. The reason for this may be issues with machine precision. With very high α_D settings, the gradient of the cost function around the minimum becomes extremely steep, so that even within a highly contracted simplex, there may be multiple different solutions. The instability of predictions at extreme rates in our opinion does not constitute a strong argument against this model. After all, human speakers can possibly produce only a limited range of speaking rates. Of course, the value of $\alpha_D = 11$ is not a hard boundary in this respect, but depends on the settings of the other parameters. It may be stated with regard to our key result, the incompressibility bifurcation that it is borne out within the range of meaningful parameter settings.

We think that these are very interesting results, but unfortunately, the modified model makes a clearly incorrect prediction with regard to prosodic prominence. This is shown in Figure 7.11, plotting simulation results with increasing speaking rate and a categorical stress distinction: $\psi_i = 2$ (stressed) and $\psi_i = 4$ (unstressed: the modified model does predict higher stressed than unstressed intercepts at slow rates, but the trend is reversed at faster rates, contrary to the real data, where the stressed-unstressed difference tends to become *more* marked at the faster rate (cf. Table 7.2). This may not be overly problematic; the observation of greater intercepts for stressed than for unstressed syllables in the BTC data is somewhat preliminary because factors such as segmental differences are not controlled for the stressed-unstressed comparison. What is definitively unsettling, however, is that the model predicts stressed syllables to eventually become shorter than unstressed syllables at very fast rates (as is evident from the stressed distribution lying more closely to the y-axis than the unstressed distribution at very fast rates), and stressed syllables are also first to be deleted.

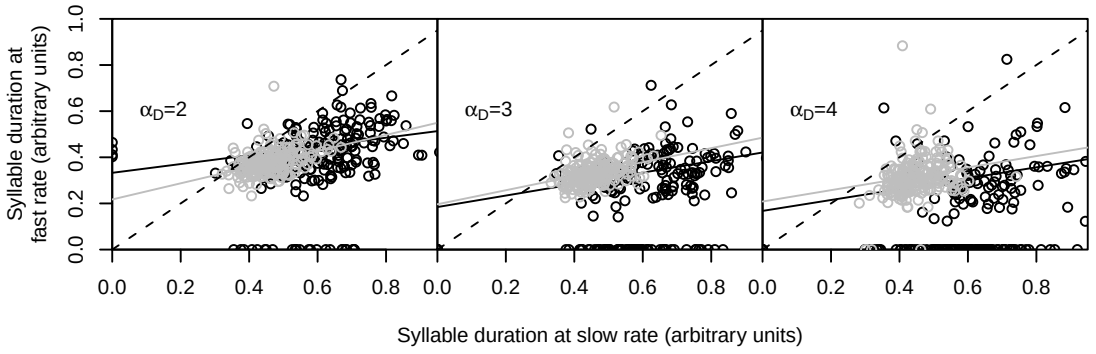


FIGURE 7.11: Simulated syllables durations at increasingly fast rates (α_D shown in plot panels) as a function of simulated syllable durations at slow ($\alpha_D = 1$) rate. Black: stressed syllables; gray: unstressed syllables. See text for details.

The cause of this incorrect prediction is evident from Figure 7.12, which shows plots of cost function C for a stressed and an unstressed syllable, with circles marking their respective minima and, hence, optimal durations. Solid lines show the cost functions at a slow rate ($\alpha_D = 1$), and dashed lines show the respective functions at a fast rate ($\alpha_D = 4$). As can be seen, the minimum of C at the slow rate is not only further to the right, but also considerably shallower for a stressed than for an unstressed syllable. This means that if overall speaking rate is increased, the stressed minimum will cross the deletion threshold earlier, as shown by the dashed trajectories: the optimal duration of the stressed syllable (marked by a black X) has gone all the way down to zero, whereas the unstressed optimal duration (gray cross) is still considerably greater.

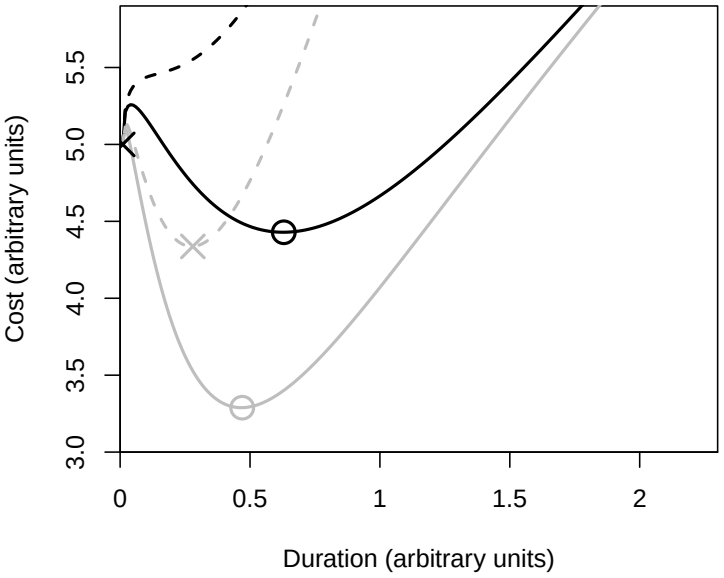


FIGURE 7.12: Plot of cost function C for a stressed (black; $\psi_i = 2$) and an unstressed (gray; $\psi_i = 4$) syllable at $\alpha_D = 1$ (solid lines) and $\alpha_D = 4$ (dashed lines). Other parameter settings are as in the simulations reported above. Circles mark optimal durations at the slow rate, X's mark optimal durations at the fast rate.

We experimented with introducing an additional assumption, using the same parameter ψ_i, for locally weighting E and P_S, so that lexical stress is modeled not only by boosting the perceptual, but also by lowering the effort-related component. This may not be a completely ad-hoc strategy; one may argue that it removes an excessive degree of freedom from the model, since independent η_i and ψ_i parameters arguably allow for simultaneously trigger hyper- and hypoarticulation of the same syllable by assigning a high weight to both parameters, which is paradoxical. However, even this additional assumption was found to rectify the incorrect prediction regarding prominence and speaking rate only under certain numerical parameter settings, and thus did not prove to be a viable solution. We have to leave it to further research to find a model that both allows for deletions and correctly predicts the influence of prosodic prominence on the observed incompressibility. Results of modified model simulations reported here shall suffice as an initial demonstration, highlighting the potential of the chosen approach.

7.4 Discussion

We have attested incompressibility of syllable durations as a function of speaking rate variation in several languages, using a straightforward regression method. Simulation experiments with the basic model as defined in Chapter 5 have demonstrated that the basic pattern of results observed in the empirical data emerges automatically from the formalization of H&H assumptions in our optimization-based model of speech timing. A

modified version of the model was shown to provide an interesting account of observed deletion phenomena, but this model proved unable to capture the influence of prosodic prominence correctly. As for the basic model, the incompressibility pattern is a rather obvious consequence of its architecture, but as we said above, the crucial point is that the responsible modeling assumptions are motivated by independent evidence. In the subsequent chapter of this work, we will see that more interesting results related to high-level linguistic structure fall out from the same model architecture.

To our knowledge, we have presented the first computationally explicit model to offer an explanatory account of durational incompressibility in speech. The initial mathematical model by Klatt (1973) was purely descriptive and did not feature any explanatory mechanisms – if anything, it utilizes incompressibility itself as an explanatory device. Similarly, the explanatory model by Katz (2010) does feature incompressibility, but it is included as an explicit assumption, rather than an emergent result of the modeling paradigm. We would therefore argue that our model supplies an important contribution to the understanding of incompressibility as a property of speech timing.

Chapter 8

Modeling Effects of Prominence, Position and External Conditions on Suprasegmental Speech Timing

8.1 Introduction

In this chapter, we will, finally, turn to the suprasegmental speech timing phenomena reviewed in Chapter 3 and address them within our optimization-based model of speech timing. We have treated incompressibility separately from these effects, because in our opinion, it represents a more basic property of speech production. By contrast, the effects we will address in this section mostly stem from high-level linguistic structure, in particular prosodic prominence.

In the simulations reported in this chapter, we will continue to use the basic, "non-deletion" model as introduced in Chapter 5. Yet, the modified model used in the second section of the previous chapter makes the same predictions for most of the effects to be reported here, and elsewhere (Windmann et al. 2015b), we report results of these experiments with the modified model architecture. We stick with the basic model here as a more conservative approach. Cases where the predictions (or explanations) of the two models diverge will be discussed. The structure of this chapter will differ somewhat from the order in which we introduced the difference classes of effects in Chapter 3: we will discuss positional effects and their interactions last, because our treatment of these

effects will be somewhat more speculative than the other timing phenomena treated in this work.

Unless noted otherwise, all simulations reported in this section were run on the "utterance" depicted in Figure 8.1 (re-plotted from Figure 7.3.1), i.e. a sequence of eight syllables, with the first, the fourth and the penultimate syllable being stressed ($\psi_i = 1$), all others unstressed ($\psi_i = 0.5$), and the fourth and fifth syllable forming an accented "word" ($\Psi_j = 2$ for this word and 0 elsewhere; $\alpha_{PW} = 1$). Global trade-off parameters were generally set to $\alpha_E = 3$, $\alpha_P = 1$ and $\alpha_D = 1$ (as in the "slow" simulation reported in Chapter 7). The values of the local parameters η_i and δ_S were set to 1. Different parameter settings will be indicated wherever used. Experimentation confirmed that results to be reported are not affected by modifications of the input utterance, such as adding or removing syllables, or changing the number or distribution of stresses. The exception to this is the number of syllables in the accented "word", as will be reported in the experiment on constituent length effects below.

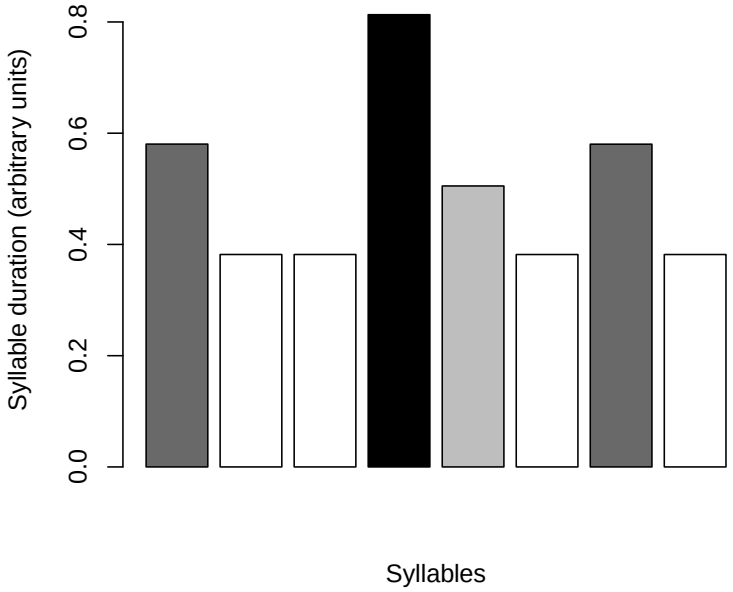


FIGURE 8.1: Syllable durations predicted by the model for a hypothetical utterance with a bisyllabic accented word. Black: +stress +accent; light gray: -stress +accent; dark gray: +stress -accent; white: -stress -accent (re-plotted from Figure 7.3.1).

8.2 Prominence Effects

8.2.1 Interaction of Stress and Accent

As we have reported in Chapter 3, evidence from various language suggests that accentual lengthening affects all syllables within a word, or at least within some domain

that extends beyond the accent-bearing syllable itself. It was shown that within this domain, lengthening is typically not uniformly distributed: there seems to be a cross-linguistic tendency for stressed vowel durations to increase by a greater proportion than unstressed vowel durations under accentual lengthening. A potential caveat was that the difference may be neutralized in word-final position. We will ignore this complication for the time being and concentrate on the basic durational pattern, i.e., stronger lengthening of stressed than unstressed vowels.

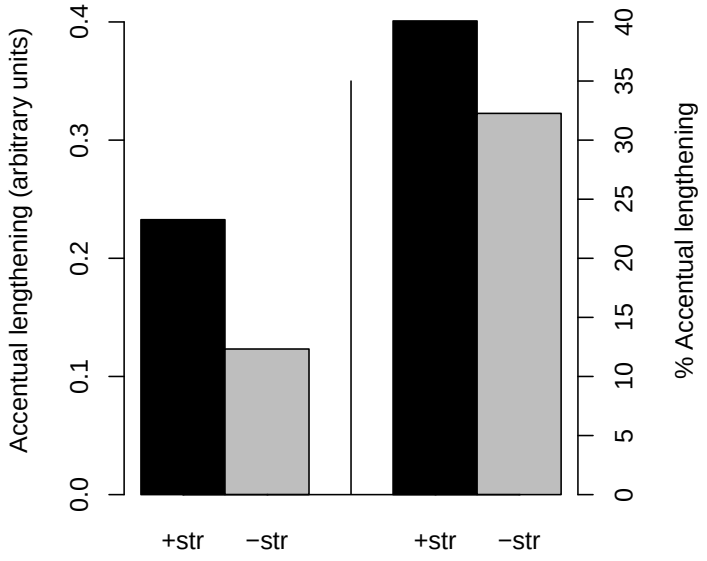


FIGURE 8.2: Absolute (left) and proportional (right) amount of accentual lengthening in stressed and unstressed syllables in the simulated utterance (bisyllabic accented word).

The model was run twice on the utterance depicted in Figure 8.1, i.e., an utterance with a bisyllabic accented word, with the above parameter settings. Accentual lengthening in the *stressed* syllable was defined as the absolute and percentage difference in duration between the fourth (stressed accented) and the first (stressed unaccented) syllable in the utterance in Figure 8.1; accentual lengthening in the *unstressed* syllable was defined as the absolute and percentage difference in duration between the fifth (unstressed accented) and the second (unstressed unaccented) syllable in the utterance in Figure 8.1 (as is apparent from Figure 8.1, the choice of the unaccented reference syllables is arbitrary, as unaccented syllables sharing the same stress value always have the same duration under the chosen parameter settings). Figure 8.2 shows that the model correctly predicts the qualitative pattern of results reported in the literature: the effect of accentuation is greater in the stressed than in the unstressed syllable, in absolute as well as proportional terms.

Experimentation with the model showed that the proportionally greater lengthening of the stressed syllable crucially depends on the nonlinearity of E: if the square root here is

exchanged for a linear function, proportional accentual lengthening is exactly the same in stressed and unstressed syllables. The mathematical explanation is fairly obvious: the concave nonlinearity of the square root function makes lengthening an already long syllable cheaper than lengthening a shorter syllable by the same amount of duration. It may seem at first glance that this tendency is counteracted by the mathematical properties of P_S, whose convex nonlinearity would appear to suggest that lengthening short syllables will generate relatively more benefit than lengthening already longer ones. It must be kept in mind, however, that as for P_S, stressed and unstressed syllables actually lie on different lines, due to the multiplicative stress parameter ψ_i.

A critical observer may point out that using the square root for E does in fact little more than explicitly instructing the model to lengthen stressed syllables more than unstressed ones in accented words. To this we would reply, again, that the concave shape of E is not arbitrary; we motivated it based on our mass-spring model simulations in Chapter 3. The explanation suggested by our model is truly related to production-perception trade-offs: accentuation, on this account, is interpreted as a perception-driven impetus to lengthen a word, so as to make it more prominent. The distribution of this lengthening among the component syllables of the word is realized in an efficient fashion, namely so as to incur the lowest possibly increase in effort.[1]

With a view to the particular motivation of cost function E as sketched in Chapter 5, we may develop this explanation a bit further. We motivated the concave non-linearity of E based on the reasoning that at the lower end of the temporal scale, both articulation and phonation contribute to the effort required to produce a syllable, whereas for longer durations where articulatory targets have eventually been reached, further lengthening only requires sustained phonation. Applying this reasoning to the stress-accent interaction, our model suggests the following tentative explanation: accentual lengthening in stressed syllables happens in the region of the temporal scale where lengthening is mainly achieved by stretching the steady state of the vowel, as stressed vowels are already relatively long and peripheral. Unstressed vowels are shorter, and also tend to be less peripheral than stressed vowels, hence our model predicts that hyperarticulating them under the influence of accent adds effort on both counts, articulation and phonation. The most efficient solution, then, is to lengthen mainly the stressed syllable, which can be had for relatively less *articulatory* effort.

Is this explanation plausible? It is certainly not true that accentual lengthening has *no* correlates in the articulatory domain; for example, studies such as Fourakis et al.

[1] Based on results showing that accentual lengthening may "skip" syllables in multisyllabic words (e.g. Dimitrova and Turk 2012), Turk (2014) argues that accentual lengthening does not affect the word as a whole, but multiple sites within a word (which appear to form a continuous domain in short words). However, nothing about our basic result would change if P_W in the model was applied to a discontinuous sequence of syllables.

(1999) or De Jong (2004) show that stressed accented as compared to stressed unaccented vowels are characterized by more peripheral F1 values, indicating increased jaw opening. However, these changes are relatively subtle compared to the large durational effects of prosodic prominence. De Jong (2004)'s analysis of formant measurements on English vowels of increasing prominence provides a case in point: The effect of accent on F1 is subtle and inconsistent across consonantal contexts, whereas durational effects are large and reliable. This pattern is also intuitively compelling: articulatory expansion as a function of prosodic prominence faces relatively low upper bounds, dictated by the anatomical limits of the vocal tract and the necessity to secure perception of the desired vowel quality. By contrast, purely temporal hyperarticulation by means of sustaining vowel production is theoretically possible for very long time spans, exceeding the durations typically found in speech by several orders of magnitude. It remains to be investigated why, on this account, the stress-accent interaction is not observed in Spanish, as discussed in Chapter 3.

One last point to note is that the modified model presented in section 7.2 of this work, while making the same *prediction* regarding the stress-accent interaction as the basic model, suggests a different *explanation*: with the modified model, the interaction is borne out regardless of the function used for E. The cause of the stronger effect of accent on stressed compared to unstressed syllable durations in the modified model lies entirely within the different slopes of P_S, which, in the unstressed case, flattens out earlier than in the stressed case. Thus, additional lengthening caused by accent ceases to reduce perceptual cost earlier for unstressed than for stressed syllables. This explanation, thus, is entirely perception-based: successful recognition is more crucial for stressed than for unstressed syllables, hence lengthening them to a stronger extent generates more perceptual benefit.

We do not attempt to decide between the explanations of the stress-accent interaction suggested by the two versions of the model, as we know of no data that would be relevant for this decision. As a final point, the two explanations are certainly not mutually exclusive, hence we do not regard it as problematic that the two versions of the model provide two different explanatory accounts of the stress-accent interaction.

8.2.2 Constituent Length Effect in Accented Words

We have reviewed constituent length effects, i.e., putative (inverse) relationships between syllable or vowel duration and syllable count in larger prosodic domains in Chapter 3, and contributed our own empirical investigation of such effects in Chapter 6 of this work. Results from the literature review suggest that these effects only surface in very

restricted contexts, namely in words carrying highly prominent nuclear or contrastive pitch accents. Our own analysis actually failed to yield evidence for constituent length effects even in accented contexts, but as we argued in Chapter 6, this may have to do with the accent definition employed in this corpus, and does not invalidate results from controlled experimental studies. This state of matters is consistent with White (2002)'s interpretation that the effect is a mere epiphenomenon of accentual lengthening: on this view, the variation according to the number of syllables in the word arises because total accentual lengthening does not increase in words with higher syllable count and therefore has to be shared out among the individual syllables. Thus, shortening of syllables in polysyllabic accented words is interpreted as a direct consequence of word prominence, and not as an indication of durational mechanisms that impose a tendency to shorten or equalize durations of words or similar prosodic domains. As we shall see, it is precisely this interpretation of the observed shortening effect that our model embodies. In the description of our experiments, we will continue to use the term "polysyllabic shortening", in accordance with White (2002)'s specific usage of this term for a constituent length effect at the word level.

The polysyllabic shortening effect was tested in the model by running simulations with the parameter settings as indicated above while varying the number of syllables in the accented word between 1 and 4 by introducing additional syllables. Figure 8.3 graphs the durations of the stressed and the following unstressed syllable from the accented word in the simulated utterance as a function of the number of syllables in the accented word, illustrating the polysyllabic shortening effect. This prediction in particular is in close agreement with empirical findings: first, the decrease in syllable duration as a function of word length is not linear, but negatively accelerated, such that duration differences between syllables from words of increasing syllable count become progressively smaller. This pattern is ubiquitously reported in the literature on polysyllabic shortening. Second, the effect is relatively subtle in unstressed syllables, in accordance with results from reiterant productions in Dutch (Nooteboom 1972) and Swedish (Lindblom and Rapp 1975).

It is important to note that while P_W is of course responsible for the polysyllabic shortening effect, it does not trivially hardcode the pattern by explicitly shortening syllable durations according to syllable count in the word. A superficial observer might suspect this, because similar functions have been utilized in this way in *descriptive* models of speech timing (Lindblom and Rapp 1975, Nooteboom 1972). We would like to reiterate at this point that P_W in our model does, in fact, precisely the opposite: it is a *cost* on shortening the accented word, hence providing an impetus to lengthen all syllables

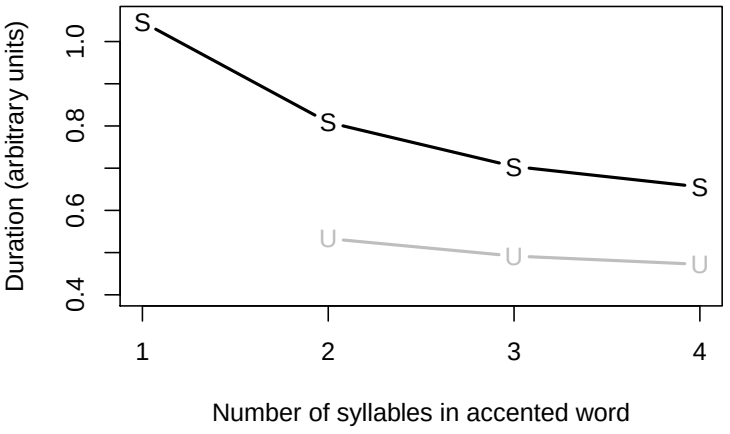


FIGURE 8.3: Polysyllabic shortening in stressed (black) and unstressed (gray) syllables in the accented word as predicted by the model. See text for details.

within it. Moreover, note that P_W, being a function of the sum of syllable durations within its scope, is blind to the number of the syllables in the accented word.

Figure 8.4 shows that accentual lengthening of the word as a whole (measured as the difference between the summed durations of all syllables in the accented "word" and the summed durations of the same number of stressed and unstressed syllables from outside the scope of the accentual lengthening) also decreases as a function of syllable count. This prediction is also borne out by empirical data reported by White (2002). Note that this is not a mere restatement of the polysyllabic shortening pattern at the syllabic level, but, in fact, an independent result: shortening of syllables as a function of syllable count in the word would also be possible with constant, and even with increasing accentual lengthening of total word duration as a function of syllable count in the word, as is easily verified by redoing the necessary calculations with appropriately constructed hypothetical durations.[2]

In order to understand how the model reproduces the polysyllabic shortening effect, it is helpful to consider total lengthening of the accented word first: accentual lengthening is sensible only as long as it still yields a reduction in overall cost, that is, if the lengthening happens within the duration region where P_W still has a a sufficiently negative gradient to outweigh the combined effect of E and D, which both assign extra costs to lengthening (which has to be outweighed by P_S and P_W). Word duration naturally increases with

[2] Consider the case of a monosyllabic (S_1), a bisyllabic (S_2U_2) and a trisyllabic word ($S_3U_3U_3$) word, where $S_1 = S_2 = S_3 = 100$ ms and $U_2 = U_3 = 50$ ms if unaccented. Assume an accented condition with $S_1 = 150$ ms, $S_2 = 130$ ms and $S_3 = 120$ ms, $U_2 = 70$ ms and $U_3 = 65$ ms. There is polysyllabic shortening in the accented condition, yet accentual lengthening of the word (i.e., the accented-unaccented difference in S_1, $\sum[S_2U_2]$, and $\sum[S_3U_3U_3]$) is always 50 ms, regardless of syllable count in the word. With the same stressed values and $U_2 = 80$ ms and $U_3 = 75$ ms, we can even construct a case where the accented-unaccented difference in total word duration *increases* with syllable count in the word (monosyllable: 50 ms; bisyllable: 60 ms; trisyllable: 70 ms) despite polysyllabic shortening.

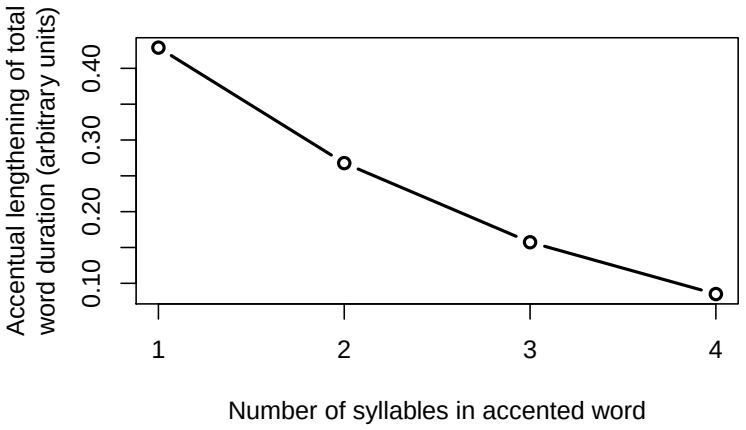


FIGURE 8.4: Accented-unaccented difference in total word duration as a function of the number of syllables in the word as predicted by the model. See text for details.

syllable count, hence the more syllables a word contains the closer it is to the point where P_S and P_W flatten out and E and D start to dominate the cost landscape and induce shortening. Consequently, total accentual lengthening will be reduced in words that contain more syllables. As for individual syllable durations, in turn, distributing the accentual lengthening among them ensures that the lengthening of each individual syllable stays within the durational range in which lengthening reduces overall cost.

The explanation of the polysyllabic shortening effect offered by our model is thus very much in keeping with White (2002)'s theoretical account of suprasegmental speech timing: polysyllabic shortening is a mere epiphenomenon of word prominence, representing the optimal distribution of the extra time in an accented word. It also offers an explanation as to why accentual lengthening does not increase in polysyllabic words, namely the hypothesized property of the perceptual cost functions in the model to eventually flatten out, as the possible gain in recognizability reaches a ceiling. This contrasts with all previous explanatory accounts of constituent length effects (except for Nooteboom (1985), which assume explicit compensatory timing relationships between syllables and larger prosodic units. Of course, we do not claim that the absence of polysyllabic shortening outside the accented word is a result explained by the model, as we explicitly set Ψ_j to 0 outside the accented word. If Ψ_j was set to a non-zero value throughout a simulated utterance, polysyllabic shortening would be observed in unaccented words as well. This would be at variance with findings by White and Turk (2010) (who do observe an effect compatible with polysyllabic shortening in unaccented words, but only if they have initial stress), but other studies, e.g. Turk and Shattuck-Hufnagel (2000) do find statistically reliable polysyllabic shortening effects in unaccented words. The crucial result that is borne out by the model, in any case, is that the effect emerges as an epiphenomenon of word prominence, without any explicit prescription to reduce

syllable durations as a function of syllable count in a word.

8.3 Effects of Overall Speaking Rate: Time Constraints and Global Hyperarticulation

To investigate how durational characteristics interact with speaking rate due to two sources, time constraints and variation along the Hypo-Hyperarticulation axis, we varied the parameters α_D (time constraints; from now on, we will simply refer to α_D manipulation as *(speaking) rate* manipulation, and to α_P manipulation as *H&H scale* manipulation. Results are shown in Figures 8.5 and 8.6. The upper panel in both figures shows the influence of the respective parameter on absolute durations. Again, the simple fact that duration decreases with increasing rate (increasing α_D) and increase (higher α_P) merely show that the model works as expected. By themselves, these patterns do not convey any theoretically interesting results. The crucial question is whether the model predicts the global parameters to affect syllables in different prosodic conditions to different degrees, which is not explicitly encoded by any parameter setting. This will be investigated by examining duration ratios between stressed and unstressed syllables in accented and unaccented contexts, and, conversely, between accented and unaccented syllables with and without lexical stress.

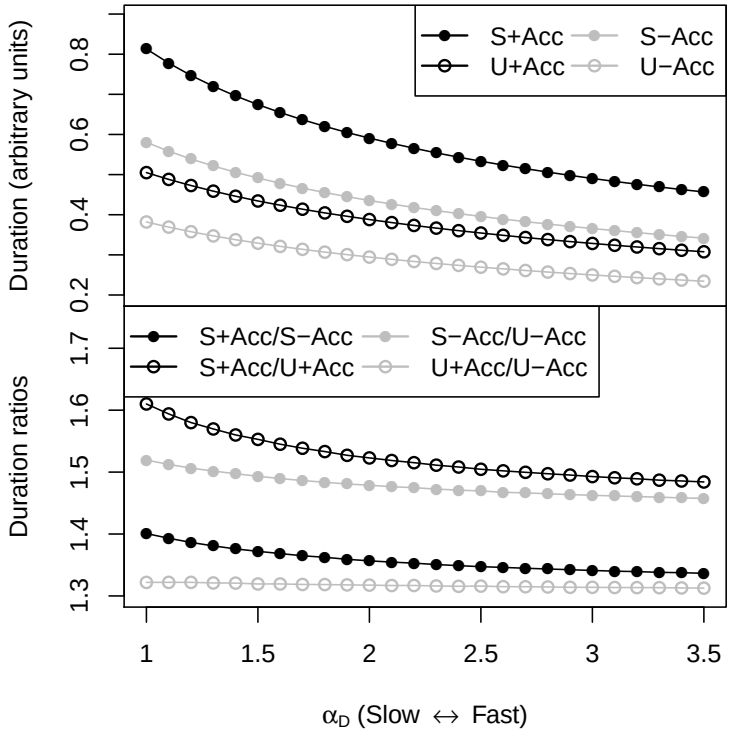


FIGURE 8.5: Influence of overall speaking rate (parameter α_D) on predicted syllable durations. Upper panel: absolute durations; lower panel: duration ratios.

Figure 8.5 shows results for variation of the rate parameter α_D, documenting the expected decrease in durations as the rate parameter is increased. As is apparent from the decaying "S+Acc/U+Acc" and "S-Acc/U-Acc" duration ratio trajectories in the lower panel, stressed syllable durations are affected more strongly by this rate variation than their unstressed counterparts. This is the case particularly in accented contexts, whereas the effect is relatively subtle in unaccented contexts. The model also predicts accentual lengthening in stressed syllables, indicated by the "S+Acc/S-Acc" ratio, to decrease in faster speech. As we have seen in Chapter 3, the prediction of greater rate sensitivity of stressed than unstressed durations is in conflict with results from Dutch reported by Janse et al. (2003), but the majority of empirical studies – Fourakis (1991) for American English, Fourakis et al. (1999) for Greek, den Os (1988) for Dutch, Pasdeloup et al. (2006) for French, and Nadeu (2014) for Spanish and Catalan – support the prediction of our model. As was argued, decreased prominence contrasts in faster speech also appear plausible given the observation that fast speech contains fewer audible prominences than slower speech (e.g. Crystal and House 1990). The prediction of stronger shortening in stressed than in unstressed syllables stems from the incompressibility property of our model, as detailed in Chapter 7: shortening already short syllables is perceptually more costly than shortening long syllables, an explanation also hinted at by Fourakis (1991).

The lower panel of Figure 8.5 indicates that, while stressed/unstressed ratios do decrease in faster speech, they asymptotically converge to lower bounds > 1, so that the stronger shortening of stressed than unstressed syllables will not go so far as to make them shorter in absolute terms than their unstressed counterparts, at least not within reasonable parameter settings. By contrast, if we were to plot results from the modified model here, we would, again, see the prediction that stressed syllables eventually become shorter than unstressed ones, as discussed in Chapter 7 . Moreover, both models diverge on the prediction regarding accentual lengthening (the "S+Acc/S-Acc" comparison), which is predicted to become *stronger* in faster speech by the modified model. The divergence, again, stems from the slightly different implementation of the prominence parameter in both versions of the model.

We discussed possible causes for the discrepancy between the results of the Janse et al. (2003) study and those of other investigations of stress and speaking rate in Chapter 3. On a final note, we may offer a speculative possibility to unify the conflicting outcomes, allowing for the assumption that Janse et al. (2003)'s result reflects the true pattern for Dutch and is not an artifact of any methodological shortcomings: in experimenting with the model, we found that if the inverse of *squared* syllable duration, ψ_i/s^2, is used for P_S, the model predicts the pattern reported by Janse et al. (2003), i.e., stronger proportional shortening of unstressed than of stressed syllables at increasing rates, while the other predictions of the model remain qualitatively unchanged. Obviously, this solution has

to remain in the realm of speculations, as there is no principled reason to assume that P_S should be conceived differently in Dutch than in other languages.

In a further series of simulations, the global perception cost parameter α_P was varied. As explained above, variation of this parameter is interpreted as global variation on the Hypo-/Hyperarticulation continuum, caused by external conditions that increase or decrease the impetus towards maximizing communicative success. Results are shown in Figure 8.6. The figure shows that, as expected, durations increase in more hyperarticulated speech. The interesting result, again, is that syllables with different levels of prominence are affected to different degrees: the model predicts that prominence contrasts actually decrease slightly in more hyperarticulated speech, at least in accented contexts ("S+Acc/U+Acc"), whereas stressed/unstressed duration ratios in unaccented contexts remain virtually stable ("S-Acc/U-Acc"). Accentual lengthening of stressed and unstressed syllables is also predicted to decrease proportionally in hyperarticulated speech, as indicated by the "S+Acc/S-Acc" and "U+Acc/U-Acc" trajectories. These predictions seem intuitively implausible, as one may expect prosodic contrasts to increase in hyperarticulated speech. Yet, we note that the most tightly controlled study of stressed/unstressed duration ratios in hyperarticulated speech, Cutler and Butterfield (1991), supports the prediction of our model, at least if it is assumed that the stressed syllables in this study were also accented. Unfortunately, the modified model predicts exactly the opposite pattern of results, i.e., enhanced prominence contrasts in more hyperarticulated speech, more in line with the studies by Fant et al. (1991b) Patel and Schell (2008) and Cho et al. (2011). Given this state of matters, we are hesitant to make strong claims about our model's ability to account for durational interactions between local and global hyperarticulation, and note that this phenomenon also requires further empirical work.

8.4 Positional Effects

In the remainder of this chapter, we will present a rather preliminary exploration of positional effects in the basic model. As we discussed earlier, the nature of position-related lengthening effects in speech is less straightforward than that of prosodic prominence, and it is therefore not completely clear how it should be incorporated in the model. We have seen in Chapter 3 that final lengthening effects at major prosodic boundaries in particular have been interpreted as instances of hypothesized biomechanical properties of motor systems, or, alternatively, as actively employed communicative signals. While we favor the second explanation, there is, to our knowledge, currently no evidence that

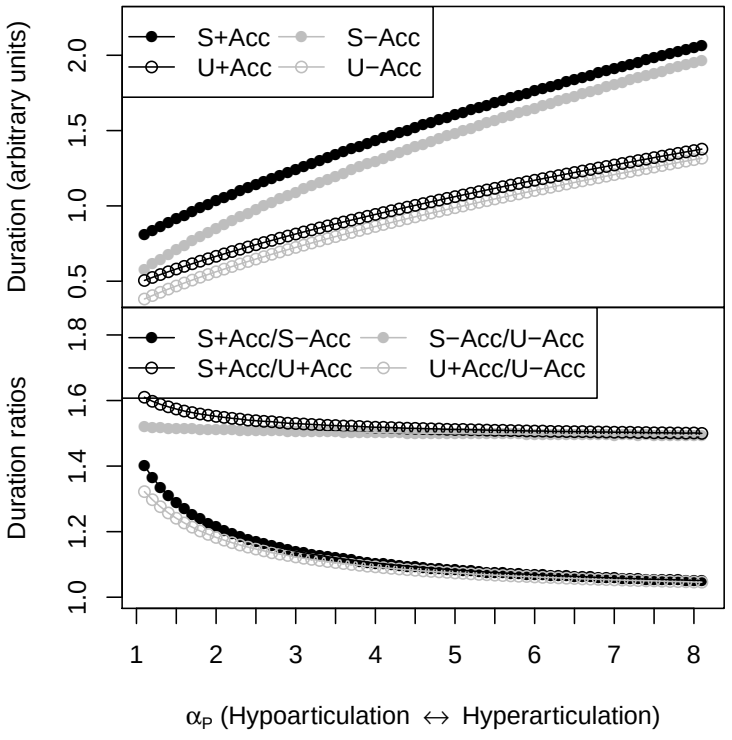


FIGURE 8.6: Influence of global variation on the H&H scale (parameter α_P) on predicted syllable durations. Upper panel: absolute durations; lower panel: duration ratios.

would conclusively settle this issue. Our treatment of positional effects will therefore be more of a tentative exploration of the model's parameter space.

Here, we present some initial results achieved with modeling positional effects by means of manipulating the local η_i parameter. There is no particularly good independent motivation as to why positional effects should be modeled in exactly this way; as we said in Chapter 5, what it conveys is that positional effects represent something of "a different kind of prominence", in that they are also characterized by a local dominance of perceptual requirements, which, however, is modeled not directly by boosting the perceptual cost function, but indirectly by weakening the relative importance of the conflicting effort-related component cost. As we shall see, this technique allows us to derive some interesting results.

We ran two simulations on the utterance described above, using the basic parameter settings as described in the beginning of this chapter without accentual lengthening, but simulating final lengthening by setting η_i to 0.1 in the final syllable of the utterance (the parameter was set to 1 elsewhere). In one of the two simulations, the final syllable was also made stressed by changing ψ_i for this syllable from 0.5 to 1. Figure 8.7 shows absolute and proportional duration changes caused by the manipulation of η_i separately for the stressed and the unstressed syllable (computed relative to any other non-accented

stressed and unstressed syllables from the utterance). The model predicts a slightly larger absolute amount of final lengthening in stressed than in unstressed syllables, but the *proportional* effect of final lengthening is stronger in the unstressed syllable. This is at variance with Campos-Astorkiza (2014)'s findings for Tuscan Italian, but roughly in line with results from American English and Hebrew discussed in Chapter 3 (Berkovits 1994, Nakatani et al. 1981, van Santen 1992)), at least as far as the proportional result is concerned. The explanation is fairly simple: as we explained earlier, it is more expensive to lengthen a short syllable by the same amount of duration than a longer syllable due to the concave nonlinearity of E. If the impact of E is locally lowered by manipulating η_i, this behavior is observed to a lesser extent.

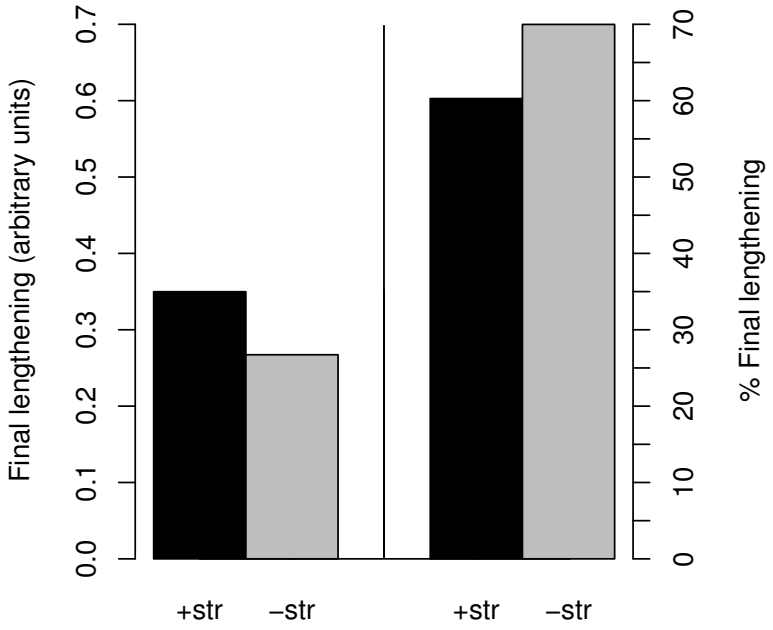


FIGURE 8.7: Absolute (left) and proportional (right) amount of final lengthening in stressed and unstressed syllables from an unaccented context in the simulated utterance.

This result also extends to phrasal prominence: we repeated the simulation with the utterance-final stressed syllable twice, this time manipulating accentual lengthening ($\Psi_j = 2$ in the accented and $\Psi_j = 0$ in the unaccented case) of this syllable, assuming that it constitutes a monosyllabic word. Results are shown in Figure 8.8. The pattern of results is similar to the stressed-unstressed comparison in the unaccented case: absolute final lengthening is slightly larger in the accented than in the unaccented monosyllable, whereas the proportional effect of final lengthening is larger in the unaccented case, as the lengthening effect of final position is applied to a comparatively smaller base duration.

In Chapter 3, we reviewed studies of accentual lengthening in minimal stress pairs and words composed of reiteranct syllables (Cambier-Langeveld and Turk 1999, Sluijter and Van Heuven 1996, Sluijter 1995), which suggest that the finding of proportionally greater

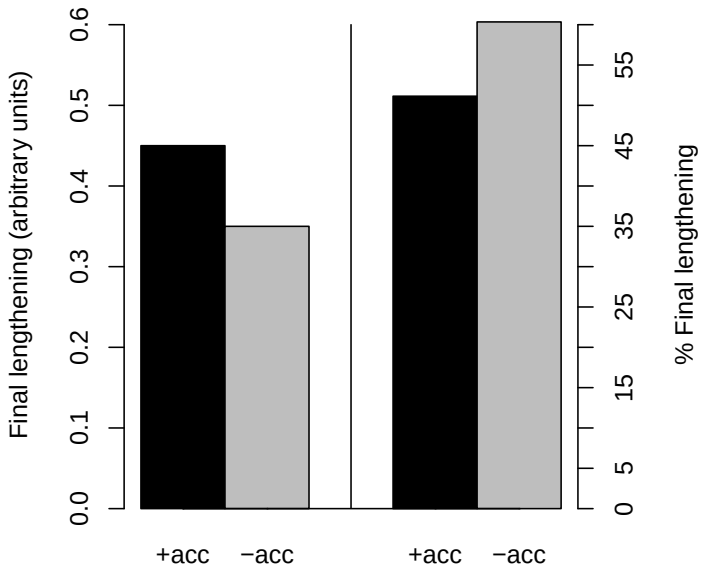


FIGURE 8.8: Absolute (left) and proportional (right) amount of final lengthening in an accented versus unaccented stressed monosyllable from the simulated utterance.

accentual lengthening in stressed compared to unstressed syllables may not hold in word-final position. As we mentioned, Sluijter and Van Heuven (1996) explicitly speculate that this may be indicative of a three-way interaction between stress, accent and word-final lengthening. We tested this hypothesis by running two simulations with a bisyllabic accented word ($\Psi_j = 2$) with either initial or final stress and η_i set to 0.5 for the word-final syllable in either case, assuming that word-final lengthening is weaker than final lengthening at larger constituent boundaries. Either simulation also included one syllable from an unaccented context that was subject to final lengthening and had the same stress level as the final syllable from the accented word. Proportional accentual lengthening estimates by stress-level (stressed/unstressed) and position (final/non-final) were obtained from comparisons of accented and unaccented syllables with the same position and stress level within the simulated utterances. Results are shown in Figure 8.9. The simulation reproduces the empirically observed pattern of results: in word-initial position (i.e., where $\eta_i = 1$), proportional accentual lengthening is stronger in stressed than in unstressed syllables. In final position, ($\eta_i = 0.5$), proportional accentual lengthening is roughly equal in stressed and unstressed syllables. The latter prediction is of course a consequence of the particular numerical parameter setting of $\eta_i = 0.5$ for word-final lengthening; with a more extreme parameter setting, accentual lengthening in word-final position would actually become stronger in unstressed than in stressed syllables. What is important, though, is that the model does replicate the general finding that the overmultiplicative durational interaction between stress and accent does not hold in word-final position. Thus, our modeling approach suggests that final lengthening

may indeed be a possible cause for the reversal of the stress-accent interaction in word-final syllables, as suggested by Sluijter and Van Heuven (1996). One caveat is that our simulation predicts generally stronger accentual lengthening in word-final than in word-initial position in both stressed and unstressed contexts, which is supported only for unstressed syllables in the cited studies. The result for stressed syllables does, however, converge with results of our own corpus analysis in Chapter 6, where we observed a stronger lengthening effect of accent in word-final than in non word-final vowels.

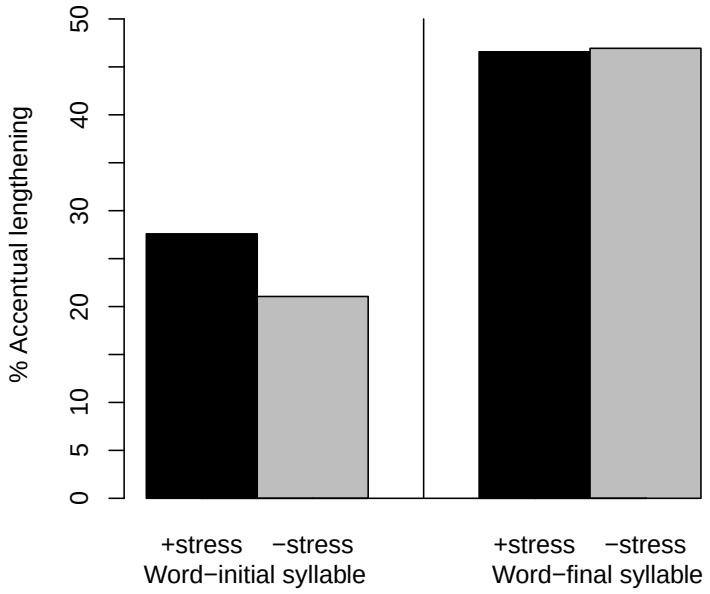


FIGURE 8.9: Proportional accentual lengthening in stressed and unstressed syllables in word-initial (left) and word-final (right) position (word-final lengthening: $\eta_i = 0.5$).

A series of simulations were run on two utterances with a stressed and an unstressed final syllable at increasing rates. Results are shown in Figure 8.10, in the same fashion as in Figures 8.5 and 8.6 above. The Figure indicates that the final lengthening effect becomes proportionally weaker under increasing rate, which may be correct for unstressed syllables (Beckman and Edwards 1990), but apparently not for stressed syllables (Smith 2002, Weismer and Ingrisano 1979). We did not investigate the investigate of the global hyperarticulation parameter α_P parameter on the strength of final lengthening, as we are not aware of conclusive empirical results on this interaction.

Moreover, we currently cannot think of any non-trivial technique that would allow the model to replicate the finding that utterance-final lengthening is progressive, starting at the last stressed syllable in the utterance, or the observation that unstressed syllables within this interval may be "skipped" by final lengthening (Turk and Shattuck-Hufnagel 2007). In any case, the approach towards incorporating final lengthening effects taken in this work has been preliminary and is to be seen as an exploration of possibilities, rather than a well-motivated attempt at truly explaining positional effects on suprasegmental

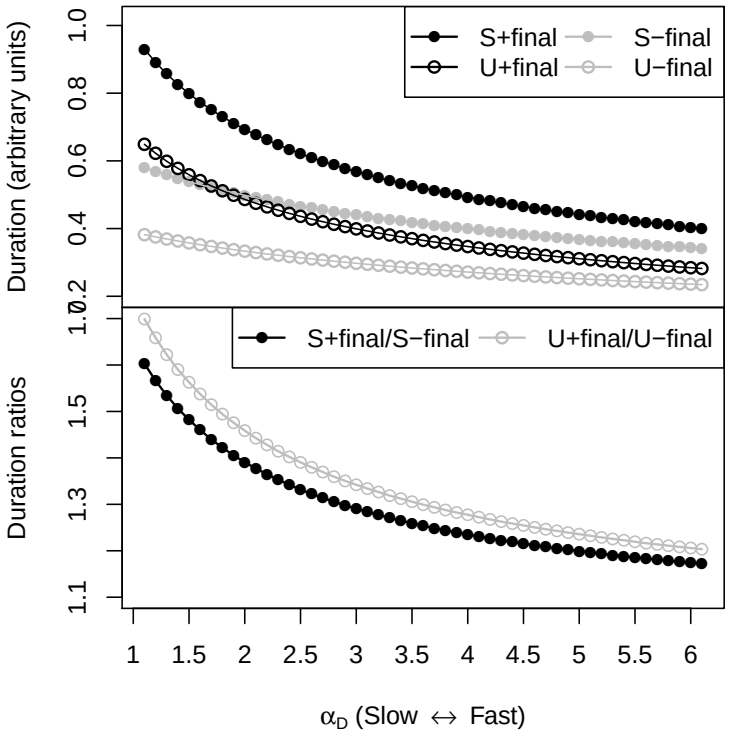


FIGURE 8.10: Influence of overall speaking rate (parameter α_D) on final lengthening in simulated stressed and unstressed syllables. Upper panel: absolute durations; lower panel: duration ratios.

speech timing. More research into the nature of these effects is clearly necessary before convincing explanatory accounts can be formulated.

8.5 Discussion

Our simulation experiments have demonstrated how several effects – or, to be more precise, *interactions* – in the domain of suprasegmental speech timing emerge automatically from the formalization of the independently motivated requirements to minimize effort and maximize communicative success in our optimization-based model. These results indicate that our model represents a promising explanatory platform for suprasegmental speech timing phenomena. This is especially true for effects and interactions involving prosodic prominence, for which the "localized hyperarticulation" account provides a well-motivated hypothesis. We are less confident about the interaction of prominence and speaking rate variation induced by global hyperarticulation. More empirical research on this topic is necessary in order to get a clearer picture of the facts.

As for position-related lengthening effects, the proposed technique has proven capable of reproducing some, but not all observed patterns. In general, our account of these

phenomena is more tentative than that of prominence, given that the causes of position-related lengthening in speech are not entirely clear. Hence, an equally well-motivated explanatory mechanism has yet to be established. Of course it has to be kept in mind that our model represents a relatively simple first pass at modeling suprasegmental speech timing phenomena and can, in this case, only serve as a generator for initial hypotheses.

We saw that the predictions and explanations suggested by the two version of our model, the "deletion" and the "baseline" approach diverge in some cases. It would be a desirable achievement for future research to find a way to unify both approaches that combines the interesting properties of the "deletion" model with the results gained with the "baseline" approach reported in this chapter.

Chapter 9

Conclusion

This work has investigated the hypothesis that speech timing patterns at the suprasegmental level result from trade-offs between the competing requirements of minimizing effort and maximizing communicative success. This assumption has been formalized in a computational model, using cost optimization to derive speech timing patterns that optimally satisfy both requirements and the additional dimension of global rate constraints. Results of simulation experiments show that the quite simple model we developed, although not successful on all counts, provides a promising approach towards accounting for suprasegmental speech timing patterns, at least in stress-accent languages. The optimization paradigm as an overarching conceptual apparatus provides a cognitively plausible modeling platform that allows for deriving principled explanations for attested speech timing patterns. This is achieved on the basis of the component cost functions of the model, whose overall shapes (though not their precise mathematical implementation details, to which no theoretical status is attached) are motivated by independent assumptions about production- and perception-related factors. Our results thus add to the body of evidence supporting the assumption that production-perception trade-offs as conceived by H&H theory are an important determinant of human communication.

To our knowledge, our model is the first to address most of the empirical phenomena discussed in this work. This is true for the simulations of speaking rate and global hyperarticulation effects, as well as for the stress-accent interaction. As discussed in Chapter 2, Incompressibility has been incorporated in the optimization model of segmental timing effects by Katz (2010), but in the form of an explicit assumption, rather than as an emergent result of the modeling paradigm. Finally, the account of constituent length effects offered by our model is the first to explicitly formalize the hypothesis that they are a by-product of word prominence. This contrasts with most other explanatory accounts of suprasegmental speech timing as discussed in Chapter 4 of this work, which posit

special mechanisms to account for constituent length effects. We would argue that our account of this phenomenon is more conservative, as it falls out from the independently established concept of word prominence and does not require any specifically tailored components to predict constituent length effects. Moreover, results of our own empirical investigation in Chapter 6 suggest that in English, for which most results on such effects are reported, they are actually not a common phenomenon in running speech and seem to be restricted to highly prominent words.

An encouraging feature of our model is that the explanations of the observed effects it suggests tend to converge with well-motivated research hypotheses. This is the case in particular for the incompressibility-based explanation of differential speaking rate effects in stressed and unstressed syllables, and the account of polysyllabic shortening as redistribution of accentual lengthening. It is an important purpose of computational modeling to demonstrate that theoretically conceived ideas actually work and generate empirically observed patterns once implemented and tested. In our opinion, our model fulfills this task in a promising manner for the domain of suprasegmental speech timing, at least as far as effects and interactions involving prosodic prominence are concerned.

A common feature of most of the results that our model successfully reproduces is that they all involve what has been referred to as *overmultiplicativity* (van Santen and Olive 1990): timing processes have proportionally greater effects on longer than on shorter constituents. (van Santen and Olive 1990). To put it differently, overmultiplicative interactions between several factors affecting constituent durations are characterized by resulting in a larger percentage change than would be expected from adding the percentage changes of the individual factors in isolation. Interactions of this type indeed seem to be quite frequent in speech timing, also at the segmental level: in addition to the interactions addressed by our model, examples of overmultiplicative interactions between effects on vowel duration have been reported (on English) for example for polysyllabic shortening and postvocalic voicing (Klatt 1973, Port 1981), speaking rate and postvocalic voicing (Gopal and Syrdal 1987), rate and phonological vowel quantity (but not quantity and postvocalic voicing; Port 1981) and prominence and postvocalic voicing (Davis and Van Summers 1989, De Jong 2004). The overmultiplicative interaction of polysyllabic shortening and postvocalic voicing, in fact, motivated the formulation of Klatt (1973)'s descriptive model of speech timing; the incompressible D_{min} constant in his model ensures that timing effects combine in an overmultiplicative fashion. Our modeling work suggests that the frequent finding of overmultiplicative interactions in speech timing is based on both perception- and production related factors. On the perceptual side, the hypothesized non-linearity of recognition probability as a function of duration suggests that proportionally larger changes have to be applied to greater base durations so as to create audible effects. This is complemented by our reasoning

regarding the non-linearity of effort expressed as a function of duration, as discussed in Chapter 3, which, as we saw, also favors overmultiplicative interactions. On the other hand, we saw for the case of final lengthening phenomena that local variation in effort requirements may account for the absence of overmultiplicativity in some interactions in speech timing, although this explanation is of course speculative and may not generalize to other situations not characterized by overmultiplicativity.

An obvious question for any computational model concerns the robustness of results under different parameter settings.The parameter space of the model is obviously infinite and there may well be settings at which some of the effects reported here will not be borne out. This is, for example likely to be the case at extremely fast settings, where many syllables are deleted and predictions become essentially meaningless, as we saw in Chapter 7. In any case, we would argue that the existence of a region in the parameter space in which various effects are reproduced is interesting enough as a result. After all, only a limited range of settings is plausible, or, indeed, physically possible, in actual human speech production – human speakers cannot talk infinitely fast, or sustain the production of a sound forever. Our modified model presented in Chapter 7 represents a simple attempt at introducing actual lower boundaries. It is obvious that a considerably more complex model incorporating explicit physiological assumptions would be needed to exclude predictions that lie outside the space of physically possible speech production.

The model we presented in this thesis is obviously highly abstract and simplified, especially in that it narrowly concentrates on timing in the acoustic domain and ignores other aspects of speech prosody. In particular, a more realistic model would have to account for the inherently multidimensional nature of prosodic prominence, affecting other prosodic parameters besides duration (Fant and Kruckenberg 1989, Heuft et al. 2000, Streefkerk 2002). The simplicity of the model is intentional, since we believe it to be a prerequisite for understanding basic processes, before more complex issues can be addressed. Nevertheless, being able to present a more fully-fledged account in the future would be a desirable achievement, and in the remainder of this Chapter, we shall sketch some possible leads towards this goal.

In an attempt at providing a model that is somewhat less abstract and removed from actual speech production than the present one, we experimented with an approach where optimization was applied directly to the computational mass-spring model used to derive the articulatory effort estimate described in Chapter 5. In this approach, the single critically-damped mass-spring system is interpreted as embodying the quasi-cyclical movement of the jaw in speech production. This model is based on the assumption that jaw cycles can be used as a coarse representation of syllables, and that prosodic structure affects these jaw movements (Erickson 1998, Lindblom 1967, Rhardisse and Abry 1995).

The model can thus be viewed as an extension of the dynamical model presented by Lindblom (1967), with the optimization component added to it: the dynamical parameters of the model, such as stiffness or target attainment, are not hand-set, but themselves the result of optimization. Compared to the present modeling paradigm, this approach adds a "vertical" perspective, as not only effects in the durational domain, but also on movement amplitudes can be studied. In our preliminary simulations, the terms of the overall cost function were defined as in ETD, using the force integral to measure effort, combining vertical target undershoot and the inverse of cycle duration into a measure of perceptual cost, and introducing a temporal cost term proportional to the duration of a whole sequence.

We were able to obtain some preliminary results using the jaw model. Supplying different vertical targets for jaw cycles, we found that the model reproduced the correlation between degree of opening and duration reported by Lindblom (1967). Increasing local prominence by boosting the perception-related cost function for a single jaw cycle also reproduced the observations of greater duration and movement amplitude, as well as lower stiffness (Kelso et al. 1985). All these findings are not hugely surprising given the explicit parameter settings, but they suggest that the model may be a promising candidate for observing more interesting effects. Unfortunately, it turned out to be non-trivial and to require additional assumptions to secure convergence of the model, and it also makes some incorrect predictions – for example, Kelso et al. (1985) report faster jaw movements for stressed compared to unstressed syllables despite the lower stiffness of the former, and this was not borne out by the model. Thus, while the "re-embodied" approach towards modeling speech timing does hold promise, more work is definitely necessary before it can be used to establish any firm conclusions.

Another possibility for providing a more realistic and comprehensive model may be to re-integrate our approach with the existing ETD platform. This could be done by directly implanting the syllabic "tier" as used in our model into ETD, defining temporal cost functions at the syllabic level for the summed activation intervals of gestures that belong to a given syllable. This approach may hold promise for modeling timing phenomena at the sub-syllabic level, such as the onset and coda shortening effects on vowel duration documented by Katz (2010), or as vowel lengthening triggered by postvocalic voicing. The latter phenomenon may be especially interesting in this regard: various explanation for the "voicing effect" have been proposed (Fowler 1992, Kluender et al. 1988). Results by Davis and Van Summers (1989) and De Jong (2004) suggest that in English, the effect may be weak or even completely absent in non-prominent contexts; (White 2014:48) also cites Klatt (1976)'s observation "that large coda voicing effects were only seen in phrase-final position". Thus, the explanation of the voicing effect may be analogous to the one for the polysyllabic shortening effect in accented words proposed by our present model:

(White 2014:ibid.) proposes that " (b)ecause voiced consonants are less expandable than voiceless consonants, the nucleus receives more final lengthening than when followed by a greatly lengthened voiceless coda". The same may apply to lengthening related to prominence. It would be interesting to test whether this hypothesis is borne out by a combination of ETD with our suprasegmental modeling paradigm. A truly satisfactory account would of course require realistic voice source modeling, but one may start out using a simplified approach, coding the difference between voiced and voiceless stops in terms of ETD's current parameter settings.[1] A foreseeable difficulty, however is that the optimization problem in ETD is already by orders of magnitude more complex than in our very simple model. Combining both approaches is likely to result in even longer computation times, and, if anything, may result in convergence issues.

A possible long-term goal for a more realistic model architecture is to simulate interaction between separate production and perception modules in an explicit production-perception loop. In this approach, the production module, informally classifiable as a "speaker" would have the task to get a message across with as little effort as possible which the "listener" would have to comprehend. Perceptual cost could be measured directly in this approach by checking whether the listener, conceivable in a similar fashion as Boersma (1998)'s perception module, has comprehended the message, possibly through a noisy channel. A genetic algorithm could be used to find parameters of the speaker's production model such that effort is minimized and communicative success is maximized. This approach would thus be akin to modeling language evolution, i.e., the hypothesized convergence of linguistic communities onto optimal speech production through repeated interaction. An somewhat similar methodology has been employed by De Boer (2000) in a computational study on the emergence of vowel systems. In contrast to De Boer (2000)'s approach, the model envisioned here would probably have to start from established linguistic categories in order to reduce complexity.

We conceive a fully-fledged model of this kind as featuring a complete articulatory synthesis module on the production and, as stated above, an algorithm informed by knowledge about auditory processing in humans on the perception side. This strategy holds the possibility of providing a truly integrated account of speech, acknowledging its inherent multidimensionality rather than narrowly focusing on a single domain, as has been done here. Implementing such a model is obviously a vastly complex task, and considerable work would have to be invested in its design. The results of the present modeling work show that even with a much simpler optimization approach, interesting insights into the nature of speech can be obtained.

[1] A coarse working solution may be to model voiced stops by locally increasing overall force via the α_E parameter, reflecting the muscular activity necessary to sustain voicing. Compare Boersma (1998)'s account of the voicing effect.

Bibliography

Abercrombie, D. (1967). *Elements of general Phonetics*. Edinburgh University Press, Edinburgh.

Adisasmito-Smith, N. and Cohn, A. C. (1996). Phonetic correlates of primary and secondary stress in Indonesian: a preliminary study. *Working papers of the Cornell phonetics laboratory*, 11:1–16.

Anbari, S. A., Włodarczak, M., and Wagner, P. (2013). Rhythmic constraints on read and rapped speech. In *Proceedings of the 14th Rhythm Production and Perception Workshop*, Birmingham.

Anderson, F. C. and Pandy, M. G. (2001). Dynamic optimization of human walking. *Journal of Biomechanical Engineering*, 123(5):381–390.

Arciuli, J., Simpson, B. S., Vogel, A. P., and Ballard, K. J. (2014). Acoustic changes in the production of lexical stress during lombard speech. *Language and Speech*, 57(2):149–162.

Arvaniti, A. (2000). The phonetics of stress in Greek. *Journal of Greek Linguistics*, 1(1):9–39.

Astésano, C., Bard, E. G., and Turk, A. (2007). Structural influences on initial accent placement in French. *Language and Speech*, 50(3):423–446.

Auran, C., Bouzon, C., and Hirst, D. (2004). The aix-marsec project: an evolutive database of spoken British English. In *Proceedings of Speech Prosody 2004*, pages 561–564, Nara, Japan.

Avanzi, M., Simon, A.-C., Goldman, J.-P., and Auchlin, A. (2010). C-prom: An annotated corpus for French prominence study. In *Proceedings of Speech Prosody 2010*, Chicago.

Aylett, M. and Turk, A. (2004). 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.

Aylett, M. and Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5):3048–3058.

Baker, R. E. and Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech*, 52(4):391–413.

Barbosa, P. (2007). From syntax to acoustic duration: A dynamical model of speech rhythm production. *Speech Communication*, 49(9):725–742.

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

Beckman, M. E. and Edwards, J. (1990). Lengthenings and shortenings the nature of prosodic constituency. In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pages 152–178. Cambridge University Press, New York.

Beňuš, Š. and Šimko, J. (2014). Emergence of prosodic boundary: continuous effects of temporal affordance on inter-gestural timing. *Journal of Phonetics*, 44:110–129.

Beňuš, Š. and Šimko, J. (2015). Prosodic boundaries in lombard speech. In *Proceedings of ICPhS 2015*, pages A–71, Glasgow.

Berkovits, R. (1991). The effect of speaking rate on evidence for utterance-final lengthening. *Phonetica*, 48(1):57–66.

Berkovits, R. (1994). Durational effects in final lengthening, gapping, and contrastive stress. *Language and Speech*, 37(3):237–250.

Boersma, P. (1998). *Functional phonology. Formalizing the interactions between articulatory and perceptual drives*. PhD thesis, University of Amsterdam.

Bolinger, D. L. (1958). A theory of pitch accent in English. *WORD – The Journal of the International Linguistic Association*, 14(2-3):1–149.

Botinis, A. (1989). *Stress and prosodic structure in Greek. A phonological, acoustic, physiological and perceptual study*. PhD thesis, Lund University.

Bouzon, C. (2004). *Rythme et structuration prosodique en anglais britannique contemporain*. PhD thesis, PhD thesis, Université de Provence.

Bouzon, C. and Hirst, D. (2004). Isochrony and prosodic structure in British English. In *Proceedings of Speech Prosody 2004*, pages 223–226, Nara, Japan.

Braun, B. and Geiselmann, S. (2011). Italian in the no-man's land between stress-timing and syllable-timing? speakers are more stress-timed than listeners. In *Proceedings of Interspeech 2011*, pages 2697–2700, Florence.

Braunschweiler, N. (1997). Integrated cues of voicing and vowel length in German: a production study. *Language and Speech*, 40(4):353–376.

Browman, C. P. and Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. In Kingston, J. and Beckman, M. F., editors, *Papers in laboratory phonology I: Between the grammar and physics of speech*, pages 341–376. Cambridge University Press, Cambridge.

Browman, C. P. and Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49(3-4):155–180.

Browman, C. P. and Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology*, 3(01):219–252.

Brunet, F. (2010). *Contributions to parametric image registration and 3D surface reconstruction*. PhD thesis, Université d'Auvergne.

Byrd, D. and Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31(2):149–180.

Cambier-Langeveld, G. M. (2000). *Temporal marking of accents and boundaries*. PhD thesis, University of Amsterdam.

Cambier-Langeveld, T. and Turk, A. (1999). A cross-linguistic study of accentual lengthening: Dutch vs. English. *Journal of Phonetics*, 27(3):255–280.

Campbell, N. (1988). Foot-level shortening in the Spoken English Corpus. In *Proceedings of the 7th FASE Symposium*, pages 489–494, Edinburgh.

Campos-Astorkiza, R. (2014). Lengthening and prosody in Tuscan Italian. *International Journal of Basque Linguistics and Philology*, XLV(1):83–109.

Chambers, J. M. and Hastie, T. (1992). *Statistical models in S*. Chapman & Hall, London.

Cho, T. and Keating, P. (2009). Effects of initial position versus prominence in English. *Journal of Phonetics*, 37(4):466–485.

Cho, T., Lee, Y., and Kim, S. (2011). Communicatively driven versus prosodically driven hyper-articulation in Korean. *Journal of Phonetics*, 39(3):344–361.

Cho, T., McQueen, J. M., and Cox, E. A. (2007). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics*, 35(2):210–243.

Chomsky, N. (1957). *Syntactic structures*. Mouton, The Hague/Paris.

Cooper, W. E., Eady, S. J., and Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America*, 77(6):2142–2156.

Crystal, T. and House, A. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *The Journal of the Acoustical Society of America*, 88:101–112.

Cummins, F. (1999). Some lengthening factors in English speech combine additively at most rates. *The Journal of the Acoustical Society of America*, 105(1):476–480.

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, 37(1):16–28.

Cummins, F. and Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, 26(2):145–171.

Cutler, A. and Butterfield, S. (1991). Word boundary cues in clear speech: A supplementary report. *Speech Communication*, 10(4):335–353.

Cutler, A. and Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, 2(3):133–142.

Cutler, A. and McQueen, J. M. (2014). How prosody is both mandatory and optional. In Caspers, J., Chen, Y., Heeren, W., Pacilly, J., Schiller, N. O., and van Zanten, E., editors, *Above and Beyond the Segments: Experimental Linguistics and Phonetics*, pages 71–82. John Benjamins, Amsterdam.

Dauer, R. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11(1):51–62.

Davis, S. and Van Summers, W. (1989). Vowel length and closure duration in word-medial VC sequences. *Journal of Phonetics*, 17(4):339–354.

De Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics*, 28(4):441–465.

De Jong, K. (2004). Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. *Journal of Phonetics*, 32(4):493–516.

De Jong, K. and Zawaydeh, B. (2002). Comparing stress, lexical focus, and segmental focus: patterns of variation in Arabic vowel duration. *Journal of Phonetics*, 30(1):53–75.

De Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America*, 97(1):491–504.

Delattre, P. (1966). A comparison of syllable length conditioning among languages. *International Review of Applied Linguistics*, 4(3):183–198.

Dellwo, V., Steiner, I., Aschenberner, B., Dankovicova, J., and Wagner, P. (2004). BonnTempo-Corpus & BonnTempo-Tools: A database for the study of speech rhythm and rate. In *Proceedings of Interspeech 2004*, pages 777–780, Jeju Island, Korea.

den Os, E. (1988). *Rhythm and tempo of Dutch and Italian: A contrastive study*. PhD thesis, University of Utrecht.

Dimitrova, S. and Turk, A. (2012). Patterns of accentual lengthening in English four-syllable words. *Journal of Phonetics*, 40(3):403–418.

d'Imperio, M. and Rosenthall, S. (1999). Phonetics and phonology of main stress in Italian. *Phonology*, 16(01):1–28.

Dogil, G. and Williams, B. (1999). The phonetic manifestation of word stress. In van der Hulst, H., editor, *Word prosodic systems in the languages of Europe*, pages 273–334. Mouton de Gruyter, Berlin/New York.

Echols, C. H. and Newport, E. L. (1992). The role of stress and position in determining first words. *Language acquisition*, 2(3):189–220.

Eddington, D. (2000). Spanish stress assignment within the analogical modeling of language. *Language*, 76:92–109.

Edwards, J., Beckman, M. E., and Fletcher, J. (1991). The articulatory kinematics of final lengthening. *the Journal of the Acoustical Society of America*, 89(1):369–382.

Erickson, D. (1998). Effects of contrastive emphasis on jaw opening. *Phonetica*, 55(3):147–169.

Eriksson, A. (1991). *Aspects of Swedish speech rhythm*. PhD thesis, University of Gothenburg.

Fant, G. and Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR*, 2:1–83.

Fant, G., Kruckenberg, A., and Nord, L. (1991a). Durational correlates of stress in Swedish, French, and English. *Journal of Phonetics*, 19(1):351–365.

Fant, G., Kruckenberg, A., and Nord, L. (1991b). Some observations on tempo and speaking style in Swedish text reading. In *Phonetics and Phonology of Speaking Styles*.

Farnetani, E. and Kori, S. (1986). Effects of syllable and word structure on segmental durations in spoken Italian. *Speech communication*, 5(1):17–34.

Farnetani, E. and Zmarich, C. (1997). Prominence patterns in Italian: an analysis of f0 and duration. In *Proceedings of the ESCA workshop on Intonation: Theory, Models and Applications*, pages 115–118.

Flege, J. and Brown Jr, W. (1982). Effects of utterance position on English speech timing. *Phonetica*, 39(6):337–357.

Flemming, E. (1997). Phonetic optimization: Compromise in speech production. *University of Maryland Working Papers in Linguistics*, 5:72–91.

Flemming, E. (2001a). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology*, 18(1):7–44.

Flemming, E. S. (2001b). *Auditory representations in phonology*. Routledge, New York.

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. In Hardcastle, W., Laver, J., and Gibbon, F., editors, *The Handbook of Phonetic Sciences, Second Edition*, pages 521–602. Wiley Online Library.

Fourakis, M. (1991). Tempo, stress, and vowel reduction in American English. *The Journal of the Acoustical Society of America*, 90:1816.

Fourakis, M., Botinis, A., and Katsaiti, M. (1999). Acoustic characteristics of Greek vowels. *Phonetica*, 56(1-2):28–43.

Fowler, C. A. (1977). *Timing control in speech production*. PhD thesis, Indiana University.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 8(1):113–133.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14(1):3–28.

Fowler, C. A. (1990). Lengthenings and the nature of prosodic constituency: comments on Beckman and Edwards's paper. In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pages 201–207. Cambridge University Press, New York.

Fowler, C. A. (1992). Vowel duration and closure duration in voiced and unvoiced stops: There are no contrast effects here. *Journal of Phonetics*, 20(1):143–165.

Friberg, A. and Sundberg, J. (1995). Time discrimination in a monotonic, isochronous sequence. *The Journal of the Acoustical Society of America*, 98(5):2524–2531.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, 1(2):126–152.

Fujimura, O. (1987). A linear model of speech timing. In Channon, R. and Shockey, L., editors, *In Honor of Ilse Lehiste*, pages 109–123. Foris, Dordrecht.

Fujimura, O. (1994). C/D model: A computational model of phonetic implementation. In Ristad, E., editor, *Language Computations. DIMACS Workshop on Human Language.*, volume 17, pages 1–20.

Fujimura, O. (2011). Temporal organization of speech utterance: A C/D model perspective. *Cadernos de Estudos Lingüísticos*, 43:9–36.

Fujimura, O. and Erickson, D. (2004). The C/D model for prosodic representation of expressive speech in English. In *Proceedings of the Fall Meeting of the Acoustical Society of Japan*, pages 271–2.

Fujisaki, H. and Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E)*, 5(4):233–242.

Gallaway, C. and Richards, B. J. (1994). *Input and interaction in language acquisition.* Cambridge University Press.

Garnier, M., Dohen, M., Loevenbruck, H., Welby, P., and Bailly, L. (2006). The lombard effect: a physiological reflex or a controlled intelligibility enhancement? In *Proceedings of the 7th International Seminar on Speech Production*, pages 255–262, Ubatuba,Brazil.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *The Journal of the Acoustical Society of America*, 63(1):223–230.

Gopal, H. (1996). Generalizability of current models of vowel duration. *Phonetica*, 53(1-2):1–32.

Gopal, H. and Syrdal, A. K. (1987). Interaction of speaking rate and postvocalic consonantal voicing on vowel duration in American English. *The Journal of the Acoustical Society of America*, 82(S1):S16–S16.

Grabe, E. and Warren, P. (1995). Stress shift: do speakers do it or do listeners hear it. In *Papers in Laboratory Phonology IV: Phonology and Phonetic Evidence*, pages 95–110. Cambridge University Press, New York.

Gray, G. W. (1942). Phonemic microtomy: The minimum duration of perceptible speech sounds. *Communications Monographs*, 9(1):75–90.

Grimm, W. (1966). Perception of segments of English-spoken consonant-vowel syllables. *The Journal of the Acoustical Society of America*, 40(6):1454–1461.

Hakokari, J., Saarni, T., Isoaho, J., and Salakoski, T. (2008). Correlation of utterance length and segmental duration in Finnish is questionable. In *Proceedings of Interspeech 2008*, pages 881–884, Brisbane, Australia.

Heldner, M. and Strangert, E. (2001). Temporal effects of focus in Swedish. *Journal of Phonetics*, 29(3):329–361.

Heuft, B., Portele, T., Wagner, P., Widera, C., and Wolters, M. (2000). Perceptual prominence. In Sendlmeier, W., editor, *Speech and Signals*, pages 97–115. Hector, Frankfurt a. M.

Hirata, Y. and Tsukada, K. (2009). Effects of speaking rate and vowel length on formant frequency displacement in Japanese. *Phonetica*, 66:129–149.

Hirst, D. (2009). The rhythm of text and the rhythm of utterances: from metrics to models. In *Proceedings of Interspeech 2009*, pages 1519–1522, Brighton.

Hirst, D. and Bouzon, C. (2005). The effect of stress and boundaries on segmental duration in a corpus of authentic speech (British English). In *Proceedings of Interspeech 2005*, pages 29–32, Lisbon.

Horne, M., Strangert, E., and Heldner, M. (1995). Prosodic boundary strength in Swedish: final lengthening and silent interval duration. In *Proceedings ICPhS*, volume 95, pages 170–173. Citeseer.

Howard, I. S. and Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. *Motor Control*, 15(1):85–117.

Hoyt, D. F. and Taylor, C. R. (1981). Gait and the energetics of locomotion in horses. *Nature*, 292:239–240.

Huggins, A. (1975). On isochrony and syntax. In Fant, G. and Tatham, M., editors, *Auditory analysis and perception of speech*, pages 456–464. Academic Press, London.

Hyman, L. M. (2006). Word-prosodic typology. *Phonology*, 23(2):225–257.

Janse, E., Nooteboom, S., and Quené, H. (2003). Word-level intelligibility of time-compressed speech: prosodic and segmental factors. *Speech Communication*, 41(2):287–301.

Jassem, W. (1952). *Intonation of Conversational English (educated Southern British)*. Number 45. Nakl. Wroclawskiego Tow. Naukowego; skl. gl.: Dom Ksiazki.

Johnson, A., Sundberg, J., and Willbrand, H. (1983). 'kölning': a study of phonation and articulation in a type of Swedish herding song. In *Proceedings of the Stockholm Music Acoustics Conference (SMAC 83)*, pages 187–202.

Jun, S. e. (2005). *Prosodic typology: The phonology of intonation and phrasing*. Oxford University Press.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In Bybee, J. and Hopper, P., editors, *Frequency and the emergence of linguistic structure*, pages 229–254. John Benjamins, Amsterdam.

Katz, J. (2010). *Compression effects, perceptual asymmetries, and the grammar of timing*. PhD thesis, Massachusetts Institute of Technology.

Keller, E. (1987). The variation of absolute and relative measures of speech activity. *Journal of Phonetics*, 15:335–347.

Kelso, J. S., Vatikiotis-Bateson, E., Saltzman, E. L., and Kay, B. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *The Journal of the Acoustical Society of America*, 77(1):266–280.

Kim, H. (2006). *Speech rhythm in American English: a corpus study*. PhD thesis, University of Illinois.

Kim, H. and Cole, J. (2005). The stress foot as a unit of planned timing: Evidence from shortening in the prosodic phrase. In *Proceedings of Interspeech 2005*, pages 2365–2368, Lisbon.

Kim, M. R. (2011). *The phonetics of stress manifestation: Segmental variation, syllable constituency and rhythm*. PhD thesis, Stony Brook University.

Kirchner, R. (1998). *An effort based approach to consonant lenition*. PhD thesis, University of California Los Angeles.

Klatt, D. (1973). Interaction between two factors that influence vowel duration. *The Journal of the Acoustical Society of America*, 54(4):1102–1104.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, 59:1208.

Klatt, D. H. (1979). Synthesis by rule of segmental durations in English sentences. In Lindblom, B. and Öhman, S., editors, *Frontiers of Speech Communication Research*, pages 287–300. Academic Press, London/New York/San Francisco.

Kleber, F. and Klipphahn, N. (2006). An acoustic investigation of secondary stress in German. *Arbeitsberichte Institut fr Phonetik Kiel*, 37:1–18.

Kluender, K. R., Diehl, R. L., and Wright, B. A. (1988). Vowel-length differences before voiced and voiceless consonants: an auditory explanation. *Journal of Phonetics*, 16:153–169.

Kochanski, G. and Shih, C. (2000). Stem-ML: language-independent prosody description. In *Proceedings of ICSLP 2000*, pages 239–242, Bejing.

Kochanski, G. and Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication*, 39(3):311–352.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press, Cambridge.

Kohler, K. (1983). Prosodic boundary signals in German. *Phonetica*, 40(2):89–134.

Krämer, M. (2009). Main stress in Italian nonce nouns. In Torck, D. and Wetzels, L., editors, *Romance Languages and Linguistic Theory 2006: Selected Papers from 'Going Romance'*, volume 303, pages 127–142. John Benjamins Publishing.

Krivokapić, J. (2013). Rhythm and convergence between speakers of American and Indian English. *Laboratory Phonology*, 4(1):39–65.

Kuehn, D. P. and Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, 4:303–320.

Kul, M. (2007). *The principle of least effort within the hierarchy of linguistic preferences: external evidence from English*. PhD thesis, Adam Mickiewicz University, Poznań.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

Ladefoged, P. (1963). Some physiological parameters in speech. *Language and Speech*, 6(3):109–119.

Lehiste, I. (1970). *Suprasegmentals*. MIT Press, Cambridge, Massachusetts.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *The Journal of the Acoustical Society of America*, 51(6B):2018–2024.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6(3):172–187.

Liljencrants, J. and Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48:839–862.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, 35(11):1773–1781.

Lindblom, B. (1967). Vowel duration and a model of lip mandible coordination. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 4:1–29.

Lindblom, B. (1968). Temporal organization of syllable production. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 9(2–3):1–5.

Lindblom, B. (1983). Economy of speech gestures. In MacNeilage, P., editor, *The Production of Speech*, pages 217–246. Springer, New York/Heidelberg/Berlin.

Lindblom, B. (1986). Phonetic universals in vowel systems. In Ohala, J. and Jeger, J., editors, *Experimental phonology*, pages 13–44. Academic Press, Orlando.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle, W. and Marchal, A., editors, *Speech production and speech modeling*, pages 403–439. Kluwer, Dordrecht.

Lindblom, B. (1999). Emergent phonology. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, volume 25.

Lindblom, B. and Engstrand, O. (1989). In what sense is speech quantal. *Journal of Phonetics*, 17(1-2):107–121.

Lindblom, B., Lubker, J., and Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics*, 7(2):147–161.

Lindblom, B., Lyberg, B., and Holmgren, K. (1981). *Durational patterns of Swedish phonology: do they reflect short-term motor memory processes?* Indiana University Linguistics Club.

Lindblom, B. and Rapp, K. (1975). Some temporal regularities of spoken Swedish. In Fant, G. and Tatham, M., editors, *Auditory analysis and perception of speech*, pages 387–396. Academic Press, London.

Löfqvist, A., Carlborg, B., and Kitzing, P. (1982). Initial validation of an indirect measure of subglottal pressure during vowels. *The Journal of the Acoustical Society of America*, 72(2):633–635.

Löfqvist, A. and Gracco, V. L. (1999). Interarticulator programming in VCV sequences: lip and tongue movements. *The Journal of the Acoustical Society of America*, 105(3):1864–1876.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon*. PhD thesis, Indiana University.

Manolescu, A., Olson, D., and Ortega-Llebaria, M. (2009). Cues to contrastive focus in Romanian. In Vigário, M., Fróta, S., and Freitas, M. J., editors, *Phonetics and Phonology: Interactions and Interrelations*, pages 71–90. John Benjamins Publishing, Lisbon.

McAuley, J. D. (1995). *Perception of time as phase: toward an adaptive-oscillator model of rhythmic pattern processing*. PhD thesis, Indiana University Bloomington, IN.

Messum, P. (2008). Embodiment, not imitation, leads to the replication of timing phenomena. In *Proceedings of Acoustics '08*, pages 2405–2410, Paris.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America*, 85(5):2114–2134.

Moers, D., Wagner, P., Möbius, B., Müllers, F., and Jauk, I. (2010). Integrating a fast speech corpus in unit selection speech synthesis: Experiments on perception, segmentation, and duration prediction. In *Proceedings of Speech Prosody 2010*, Chicago.

Moon, S.-J. and Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America*, 96(1):40–55.

Nadeu, M. (2014). Stress-and speech rate-induced vowel quality variation in Catalan and Spanish. *Journal of Phonetics*, 46:1–22.

Nakai, S., Turk, A. E., Suomi, K., Granlund, S., Ylitalo, R., and Kunnari, S. (2012). Quantity constraints on the temporal implementation of phrasal prosody in Northern Finnish. *Journal of Phonetics*, 40(6):796–807.

Nakatani, L., O'Connor, K., and Aston, C. (1981). Prosodic aspects of American English speech rhythm. *Phonetica*, 38(1-3):84–105.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Nelson, W. L., Perkell, J. S., and Westbury, J. R. (1984). Mandible movements during increasingly rapid articulations of single syllables: Preliminary observations. *The Journal of the Acoustical Society of America*, 75(3):945–951.

Nespor, M. and Vogel, I. (2007). *Prosodic phonology*. Walter de Gruyter, Berlin.

Nooteboom, S. (1972). *Production and perception of vowel duration. A study of durational properties in Dutch*. PhD thesis, University of Utrecht.

Nooteboom, S. G. (1985). A functional view of prosodic timing in speech. In *Time, Mind, and Behavior*, pages 242–252. Springer, Berlin.

Nooteboom, S. G. (1991). Some observations on the temporal organisation and rhythm of speech. In *Proceedings of ICPhS 1991*, volume 1, pages 228–237.

O'Dell, M. and Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In *Proceedings of ICPhS 1999*, pages 1075–1078, San Francisco.

O'Dell, M. and Nieminen, T. (2001). Speech rhythms as cyclical activity. In *Proceedings of the XXIst Finnish Phonetics Symposium*, pages 159–168, Turku.

O'Dell, M. and Nieminen, T. (2008). Coupled oscillator model for speech timing: overview and examples. In *Nordic Prosody: Proceedings of the Xth Conference*, pages 179–190, Helsinki.

Öhman, S. (1967). Word and sentence intonation: a quantitative model. *STL-QPSR*, 2–3:20–54.

Öhman, S. E. (1966). Coarticulation in vcv utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, 39(1):151–168.

Okobi, A. O. (2006). *Acoustic correlates of word stress in American English*. PhD thesis, Massachusetts Institute of Technology.

Oller, D. K. (1972). On the constant total amplitude theory of final syllable lengthening. *The Journal of the Acoustical Society of America*, 51(1A):102.

Ortega-Llebaria, M. and Prieto, P. (2011). Acoustic correlates of stress in Central Catalan and Castilian Spanish. *Language and Speech*, 54(1):73–97.

Pamies Bertrán, A. (1999). Prosodic typology: on the dichotomy between stress-timed and syllable-timed languages. *Language Design: Journal of Theoretical and Experimental Linguistics*, 2:103–130.

Pasdeloup, V., Espesser, R., and Faraj, M. (2006). Rate sensitivity of syllables in French: a perceptual illusion? In *Proceedings of Speech Prosody 2006*, pages 216–219, Dresden.

Patel, R. and Schell, K. W. (2008). The influence of linguistic content on the lombard effect. *Journal of Speech, Language, and Hearing Research*, 51(1):209–220.

Perneger, T. V. (1998). What's wrong with bonferroni adjustments. *BMJ: British Medical Journal*, 316(7139):1236.

Perrier, P., Payan, Y., Zandipour, M., and Perkell, J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants: a modeling study. *The Journal of the Acoustical Society of America*, 114(3):1582–1599.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1986). Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, 29(4):434–446.

Pike, K. (1945). *The Intonation of American English.* University of Michigan Press, Ann Arbor.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer, New York.

Plag, I., Kunter, G., and Schramm, M. (2011). Acoustic correlates of primary and secondary stress in North American English. *Journal of Phonetics*, 39(3):362–374.

Pluymaekers, M., Ernestus, M., and Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, 118(4):2561–2569.

Port, R. (1981). Linguistic timing factors in combination. *The Journal of the Acoustical Society of America*, 69(1):262–274.

Portele, T. (1998). Perceived prominence and acoustic parameters in American English. In *Proceedings of ICSLP'98*, pages 667–670, Sidney.

Portele, T., Heuft, B., Widera, C., Wagner, P., and Wolters, M. (2000). Perceptual prominence. In Sendlmeier, F., editor, *Speech and Signals. Aspects of Speech Synthesis and Automatic Speech Recognition. Festschrift dedicated to Wolfgang Hess on his 60th birthday*, pages 97–116. Hektor, Frankfurt a.M.

Pouplier, M. (2012). The gaits of speech: re-examining the role of articulatory effort. In Solé, M.-J. and Recasens, D., editors, *The Initiation of Sound Change: Perception, Production, and Social Factors*, pages 147–164. John Benjamins, Amsterdam.

Prieto, P., Estebas-Vilaplana, E., and del Mar Vanrell, M. (2010). The relevance of prosodic structure in tonal articulation edge effects at the prosodic word level in Catalan and Spanish. *Journal of Phonetics*, 38(4):687–705.

Prieto, P., Vanrell, M. d. M., Astruc, L., Payne, E., and Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication*, 54(6):681–702.

Prince, A. and Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science*, 275(5306):1604–1610.

Prince, A. and Smolensky, P. (2008). *Optimality Theory: Constraint interaction in Generative Grammar*. John Wiley & Sons, Malden, MA.

Prom-On, S., Xu, Y., and Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *The journal of the Acoustical Society of America*, 125(1):405–424.

Quené, H. and Port, R. (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica*, 62(1):1–13.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rhardisse, N. and Abry, C. (1995). Mandible as syllable organizer. In *Proceedings of ICPhS 1995*, volume 3, pages 556–559, Stockholm.

Rietveld, A. (1975). Untersuchung zur Vokaldauer im Deutschen. *Phonetica*, 31(3-4):248–258.

Rietveld, T., Kerkhoff, J., and Gussenhoven, C. (2004). Word prosodic structure and vowel duration in Dutch. *Journal of Phonetics*, 32(3):349–371.

Roosman, L. M. (2006). *Phonetic experiments on the word and sentence prosody of Betawi Malay and Toba Batak*. PhD thesis, Leiden University.

Rosen, K. M. (2005). Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison. *Journal of Phonetics*, 33(4):411–426.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, 1(1):43–46.

Rusaw, E. (2013). *Modeling temporal coordination in speech production using an artificial central pattern generator neural network*. PhD thesis, University of Illinois at Urbana-Champaign.

Saltzman, E., Nam, H., Krivokapic, J., and Goldstein, L. (2008). A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In *Proceedings of Speech Prosody 2008*, pages 175–184, Campinas, Brazil.

Saltzman, E. L. and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological psychology*, 1(4):333–382.

Samlowski, B., Möbius, B., and Wagner, P. (2014). Phonetic detail in German syllable pronunciation: influences of prosody and grammar. *Frontiers in Psychology*, 5(500).

Samlowski, B., Wagner, P., and Möbius, B. (2013). Effects of lexical class and lemma frequency on German homographs. In *Proceedings of Interspeech 2013*, pages 597–601, Lyon.

Selkirk, E. O. (1986). *Phonology and Syntax*. MIT Press, Cambridge, MA.

Shattuck-Hufnagel, S. and Turk, A. (2011). Durational evidence for word-based vs. prominence-based constituent structure in limerick speech. In *Proceedings of ICPhS 2011*, Hong Kong.

Siddins, J., Harrington, J., Kleber, F., and Reubold, U. (2013). The influence of accentuation and polysyllabicity on compensatory shortening in German. In *Proceedings of Interspeech 2013*, pages 1002–1006, Lyon.

Šimko, J. (2009). *The embodied modelling of gestural sequencing in speech*. PhD thesis, University College Dublin.

Šimko, J., Beňus, Š., and Vainio, M. (2014a). Hyperarticulation in lombard speech: a preliminary study. In *Proceedings of Speech Prosody 2014*, pages 869–873, Dublin.

Šimko, J. and Cummins, F. (2010). Embodied task dynamics. *Psychological review*, 117(4):1229–1246.

Šimko, J. and Cummins, F. (2011). Sequencing and optimization within an embodied task dynamic model. *Cognitive Science*, 35(3):527–562.

Šimko, J., O'Dell, M., and Vainio, M. (2014b). Emergent consonantal quantity contrast and context-dependence of gestural phasing. *Journal of Phonetics*, 44:130–151.

Sluijter, A. M. and Van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical society of America*, 100:2471.

Sluijter, A. M. C. (1995). *Phonetic correlates of stress and accent*. PhD thesis, University of Amsterdam.

Smiljanić, R. and Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *The Journal of the Acoustical Society of America*, 118(3):1677–1688.

Smith, B. (2002). Effects of speaking rate on temporal patterns of English. *Phonetica*, 59(4):232–244.

Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge university press, Cambridge.

Snow, D. (1994). Phrase-final syllable lengthening and intonation in early child speech. *Journal of Speech, Language, and Hearing Research*, 37(4):831–840.

Stevens, K. N. (1986). Models of phonetic recognition II: a feature-based model of speech recognition. In *Proceedings of the Montreal Satellite Symposium on Speech Recognition, Twelfth International Congress on Acoustics*.

Strangert, E. (1985). *Swedish speech rhythm in a cross-language perspective*. PhD thesis, University of Lund.

Streefkerk, B. M. (2002). *Prominence. Acoustic and lexical/syntactic correlates*. PhD thesis, University of Amsterdam.

Strik, H. and Boves, L. (1995). Downtrend in F_0 and P_{sb}. *Journal of Phonetics*, 23(1):203–220.

Sundberg, J. (1987). *The science of the singing voice*. Illinois University Press, DeKalb, Illinois.

Suomi, K. (2007). On the tonal and temporal domains of accent in Finnish. *Journal of Phonetics*, 35(1):40–55.

Tekieli, M. and Cullinan, W. (1979). The perception of temporally segmented vowels and consonant-vowel syllables. *Journal of Speech, Language and Hearing Research*, 22(1):103.

Tilsen, S. (2009). Multitimescale dynamical interactions between speech rhythm and gesture. *Cognitive Science*, 33(5):839–879.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature neuroscience*, 7(9):907–915.

Trouvain, J., Barry, W. J., Nielsen, C., and Andersen, O. (1998). Implications of energy declination for speech synthesis. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*.

Turk, A. (2010). Does prosodic constituency signal relative predictability? a smooth signal redundancy hypothesis. *Laboratory Phonology*, 1(2):227–262.

Turk, A. (2014). Durational effects of phrasal stress. In Caspers, J., Chen, Y., Heeren, W., Pacilly, J., Schiller, N., and van Zanten, E., editors, *Above and Beyond the Segments: Experimental linguistics and phonetics*, pages 311–322. John Benjamins Publishing Company, Amsterdam.

Turk, A. and Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, 35(4):445–472.

Turk, A. and Shattuck-Hufnagel, S. (2014a). Timing in talking: what is it used for, and how is it controlled? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1658):20130395.

Turk, A. E. and Sawusch, J. R. (1997). The domain of accentual lengthening in American English. *Journal of Phonetics*, 25(1):25–41.

Turk, A. E. and Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics*, 28(4):397–440.

Turk, A. E. and Shattuck-Hufnagel, S. (2014b). A sketch of an extrinsic timing model of speech production. In *Proceedings of Speech Prosody 2014*, pages 241–245, Dublin.

Turk, A. E. and Shattuck-Hufnagel, S. (2015). Is there a general motor basis for final lengthening? In *Proceedings of ICPhS 2015*, pages A–77, Glasgow.

Turk, A. E. and White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics*, 27(2):171–206.

Van Lancker, D., Kreiman, J., and Bolinger, D. (1988). Anticipatory lengthening. *Journal of Phonetics*, 16(3):339–347.

van Santen, J. and d'Imperio, M. (1999). Positional effects on stressed vowel duration in Standard Italian. In *Proceedings of ICPhS 1999*, pages 1757–1760, San Francisco.

van Santen, J. P. (1992). Contextual effects on vowel duration. *Speech Communication*, 11(6):513–546.

van Santen, J. P. and Olive, J. P. (1990). The analysis of contextual effects on segmental duration. *Computer Speech & Language*, 4(4):359–390.

Van Son, R. J. and Van Santen, J. P. (2005). Duration and spectral balance of intervocalic consonants: a case for efficient communication. *Speech Communication*, 47(1):100–123.

Vayra, M., Avesani, C., and Fowler, C. A. (1999). On the phonetic bases of vowel-consonant coordination in Italian: a study of stress and "compensatory shortening". In *Proc. 14th ICPhS*, pages 495–498.

Vogel, I., Bunnell, H. T., and Hoskins, S. (1995). The phonology and phonetics of the rhythm rule. In *Papers in Laboratory Phonology IV: Phonology and Phonetic Evidence*, pages 111–127. Cambridge University Press, New York.

Wagner, P. (2002). *Vorhersage und Wahrnehmung deutscher Betonungsmuster.* PhD thesis, University of Bonn.

Wagner, P. (2008). *The rhythm of language and speech: Constraints, models, metrics and applications.* Habilitation thesis, University of Bonn.

Wagner, P., Malisz, Z., Inden, B., and Wachsmuth, I. (2013). Interaction phonology – a temporal co-ordination component enabling representational alignment within a model of communication. In Wachsmuth, I., de Ruiter, J., Jaecks, P., and Kopp, S., editors, *Alignment in Communication: Towards a New Theory of Communication*, pages 109–132. John Benjamins, Amsterdam.

Weismer, G. and Ingrisano, D. (1979). Phrase-level timing patterns in English: effects of emphatic stress location and speaking rate. *Journal of Speech, Language, and Hearing Research*, 22(3):516–533.

Wenk, B. J. and Wioland, F. (1982). Is French really syllable-timed? *Journal of Phonetics*, 10:193–216.

White, L. (2002). *English speech timing: a domain and locus approach.* PhD thesis, University of Edinburgh.

White, L. (2014). Communicative function and prosodic form in speech timing. *Speech Communication*, 63:38–54.

White, L. and Turk, A. E. (2010). English words on the procrustean bed: polysyllabic shortening reconsidered. *Journal of Phonetics*, 38(3):459–471.

Widera, C., Portele, T., and Wolters, M. (1997). Prediction of word prominence. In *Proceedings of Eurospeech 1997*, pages 999–1002, Rhodes, Greece.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, 91(3):1707–1717.

Williams, B. and Hiller, S. M. (1994). The question of randomness in English foot timing: A control experiment. *Journal of Phonetics*, 22:423–439.

Windmann, A., Šimko, J., and Wagner, P. (2014a). Probing theories of speech timing using optimization modeling. In *Proceedings of Speech Prosody 2014*, pages 346–350, Dublin.

Windmann, A., Šimko, J., and Wagner, P. (2014b). A unified account of prominence effects in an optimization-based model of speech timing. In *Proceedings of Interspeech 2014*, pages 159–163, Singapore.

Windmann, A., Šimko, J., and Wagner, P. (2015a). Polysyllabic shortening and word-final lengthening in English. In *Proceedings of Interspeech 2015*, pages 36–40, Dresden.

Windmann, A., Šimko, J., Wrede, B., and Wagner, P. (2013). Modeling durational incompressibility. In *Proceedings of Interspeech 2013*, pages 1375–1379, Lyon.

Windmann, A., Šimko, J., and Wagner, P. (2015b). Optimization-based modeling of speech timing. *Speech Communication*, 74:76–92.

Włodarczak, M. (2014). *Temporal entrainment in overlapping speech.* PhD thesis, Bielefeld University.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics*, 27(1):55–105.

Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics*, 38(3):329–336.

Zipf, G. K. (1935). *The psycho-biology of language.* Houghton, Mifflin, Boston.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Addison-Wesley Press, Cambridge, Massachusetts.

Appendix A: Source Code of the Optimization-Based Model of Speech Timing

```
##Optimization-based model of speech timing, coded in R
##Andreas Windmann, 10/22/2015
##This model predicts syllable durations for hypothetical speech utterances,
##represented as a string of syllables. This works as follows: the vector s
##of syllable durations in the utterance is computed such that it
##minimizes the cost function C (called 'cost' here), using the built-in
##optimization function optim(). C/cost consists of components that are
##themselves functions of s or parts thereof. These functions represent
##hypothesized production- and perception related influences on speech
##timing. The details are in Windmann, A. 2015. 'Optimization-Based Modeling
##of Suprasegmental Speech Timing'. PhD Thesis, Bielefeld University.


#########################################################################
##Specification of the input utterance (if you want to change it, this
##has to be done directly in the code):

#number of syllables in the utterance to be simulated:
nsyl=8

#which syllables are stressed:
stresspos=c(1,4,7)

#which syllables form the accented word:
```

```
accwd=seq(4,5)

##Model parameter settings (if you want to change them, this
##has to be done directly in the code):

#global weighting factor alpha_e for effort-related cost function E:
alpha_e=3

#local weighting factor eta (cost function E) for utterance-medial syllables
eta_i_nf=1

#local weighting factor eta (cost function E) for utterance-final syllables
eta_i_f=0.1

#global weighting factor alpha_p for perception-related cost function
#P_s (syllable level):
alpha_p=1

#value of stress parameter psi_i for unstressed syllables:
psi_i_s=1

#value of stress parameter psi_i for stressed syllables:
psi_i_u=0.5

#global weighting factor alpha_pw for perception-related cost function P_w
#(word level):
alpha_pw=1

#local weighting factor Psi_j for word accent:
Psi_j=2

#global weighting factor alpha_d for speaking rate related cost function D:
alpha_d=1

#global weighting factor delta for speaking rate related cost function D
#(currently not used):
delta=1

##############################################################################
```

```
#initialize string of syllables:
s=rep(1,nsyl)

#construct stress vector:
psi_i=rep(psi_i_u,nsyl)
psi_i[stresspos]=psi_i_s

#construct eta vector (for final lengthening; only the last syllable is
#lengthened here):
eta_i=rep(eta_i_nf,nsyl)
eta_i[length(eta_i)]=eta_i_f

#construct delta vector:
delta_i=rep(delta,nsyl)

#store overall cost throughout optimization:
costs<<-c()

#cost computation:
fun=function(s) {

        #E: sum of square roots of syllable durations
        e=sum(eta_i*sqrt(abs(s)))

        #P_s/P_w: sum of reciprocal of syllable durations/summed syllable
        #durations within accented word
        p=sum(psi_i/(abs(s)))
        pw=max(Psi_j/sum(abs(s[accwd])))

        #Alternative versions of P_s/P_w used in 'deletion' model:
        #(psi_i_s has to be < psi_i_u in this case)
        #p=sum(exp(-psi_i*abs(s)))
        #pw=exp(-Psi_j*sum(abs(s[accwd])))

        #D: sum over all syllable durations
        d=sum(delta*abs(s))

        #compute costs (using the name 'cost' here rather than 'c' because
```

```
        #'c' is the name of a data structure in R)
        cost=alpha_e*e+alpha_p*p+alpha_pw*pw+alpha_d*d
        costs<<-c(costs,cost)
        fun=alpha_e*e+alpha_p*p+alpha_pw*pw+alpha_d*d

}

#call optimization:
solution1=optim(s,fun)
vals=solution1$par

#...a couple of times:
for (a in 1:200) {

        sol=optim(vals,fun)
        vals=sol$par

        #introduce some random noise to prevent optimization
        #from 'getting stuck' in local minima:
        vals=jitter(vals,0.005)

        }

#optimal solution:
solution=optim(vals,fun)

#predicted durations:
solution$par=abs(solution$par)

par(mfrow=c(1,2))
barplot(solution$par)

#plot evolution of overall cost over optimization runs:
plot(costs,type='l')
```

Appendix B: Source Code of the Mass-Spring Model

```
function [E,t_all,y_all,durs] = jaw1(st,Ks,Ts,targets,fig)

% Computational mass-spring model, written in matlab by Juraj Simko
% computes the trajectory of a single critically damped spring, using
% matlab's ode45 solver
% articulatory effort is computed as the force integral, i.e., the sum of
% the forces acting upon the spring over time
% Since mass is just a constant, force is proportional to acceleration here.
%
% input parameters: vertical starting position (st), vertical target vector
% (targets), stiffness vector (Ks), activation interval vector (Ts),
% plotting yes (fig=1) or no (otherwise)
% output parameters: articulatory effort (e), positions vector (y_all),
% time vector (t_all), gesture durations (durs)

% specify options for differential equation solver:
options = odeset('RelTol',1e-6,'AbsTol',1e-6, 'Events', @switch_off);

t_all = [];
y_all = [];

tp=1;

t_start=0;
y_0 = [st;0;0];
% compute gestures
```

```
for gg = 1:length(Ks),
    %opening gestures

    tspan = t_start + [0 Ts(gg)];

    ode = @(t,y)two_gest(t,y,Ks(gg),targets(gg));
    % %call differential equation solver

    [t,y] = ode45(ode,tspan,y_0,options);
    t_all = [t_all;t];
    y_all = [y_all;y];
    ind_1 = length(t_all);
    t_start = t(end);
    y_0 = y_all(end,:) ;

    %closing gestures  - same stiffness and time interval as opening

    tspan = t_start + [0 Ts(gg)];

    ode = @(t,y)two_gest(t,y,Ks(gg),0);

    [t,y] = ode45(ode,tspan,y_0,options);
    t_all = [t_all;t];
    y_all = [y_all;y];
    ind_1 = length(t_all);
    t_start = t(end);
    y_0 = y_all(end,:);

    tp=[tp length(t_all)];

end

durs=diff(t_all(tp));

% % plot gesture trajectory and velocity profile:
if fig==1,

figure
subplot(3,1,1)
```

```
 plot(t_all,y_all(:,1),'b')
subplot(3,1,2)
 plot(t_all,abs(y_all(:,2)),'b')
 subplot(3,1,3)
 plot(t_all,abs(y_all(:,3)),'b')

end

% Compute effort:
E = y_all(end) - y_all(1);

end

function [dydt] = two_gest(t,y,Ks,Ts)

% specification of differential equation

Bs = 2*sqrt(Ks);
%Bs: damping coefficients
%Ks: stiffness coefficients
%  Bs = 0;
% Bs=0 -> simple harmonic oscillator
% Bs = 2*sqrt(Ks) -> critical damping

dydt = [0;0];

dydt(1) = y(2)  ;

dydt(2) = -Ks*(y(1)-Ts) - Bs(:)*y(2);

dydt(3) = abs(Ks*(y(1)-Ts) + Bs(:)*y(2));

end

function [value,isterminal,direction] = switch_off(t,y,st)

value = y(1)-0.1;

isterminal = 1;
```

```
direction = -1;


end
```