UNIVERSITY OF CALIFORNIA

Los Angeles

Maxent Harmonic Grammars and Phonetic Duration

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Linguistics

by

Lee Michael Lefkowitz

2017

ABSTRACT OF THE DISSERTATION

Maxent Harmonic Grammars and Phonetic Duration

by

Lee Michael Lefkowitz

Doctor of Philosophy in Linguistics

University of California, Los Angeles, 2017

Professor Bruce P. Hayes, Chair

Research in phonetics has established the grammatical status of gradient phonetic patterns in language, suggesting that there is a component of the grammar that governs systematic relationships between discrete phonological representations and gradiently continuous acoustic or articulatory phonetic representations. This dissertation joins several recent works in proposing that these relationships can be represented with constraint grammars, but moves from the harmonic grammars used in previous work to maxent grammars, already in common use by phonologists, describing how these can be adapted to the phonetic realm. Unlike existing models, maxent grammars allow phonetic variation to be modeled explicitly, outputting probability distributions over the realizations of phonetic variables instead of single values. The maxent formalism is shown to make a number of interesting empirical predictions regarding phonetic variation, defining a restrictive typology of possible phonetic patterns.

As a substantial case study, a grammar of this sort is developed for phonetic duration. Duration is known to be subject to a very large variety of (often conflicting) phonetic and phonological effects, and so this empirical domain is a rich testbed for theoretical research. A review of the empirical literature on duration is conducted, and a surprising generalization with regards to how effects on duration interact is discovered.

A production experiment on front vowel duration in English is conducted in order to shed light on how duration is computed by the grammar when multiple duration-related process are at play. The results replicate some of the interaction effects found by prior authors, and are remarkably consistent with the empirical predictions of the maxent framework in a number of respects.

Finally, a maxent learning algorithm for estimating the weights and the targets of phonetic constraints is described and implemented in Python, and this algorithm is trained, using several different constraint sets, on the data from the production experiment, yielding grammar fragments for English front vowel duration.

These endeavors serve, on the empirical side, as a new investigation of the how factors affecting duration interact and how they should be modeled, and on the theoretical side as an exploration of how maxent grammars behave when they are used to model continuous phonetic variables, uncovering a powerful new tool for generative phonetics.

The dissertation of Lee Michael Lefkowitz is approved.

Robert T. Daland

Sun-Ah Jun

Jody E. Kreiman

Bruce P. Hayes, Committee Chair

University of California, Los Angeles

2017

TABLE OF CONTENTS

Abstract of the Dissertation ii
Table of Contents v
List of Figures viii
List of Tables xi
Acknowledgements xiv
Vita xvi
1. Introduction 1
1.1. The case for phonetic grammars 1
1.2. Duration as a testbed 3
1.3. Roadmap of the dissertation 4
2. Duration as a phonetic variable 6
2.1. Duration defined 6
2.2. Factors affecting duration in English 9
2.2.1. Segmental factors 9
2.2.2. Word-internal prosodic factors 11
2.2.3. Phrase-level prosodic factors 14
2.2.4. Lexically specific and discourse-level factors 14
2.2.5. Speech rate 17
2.3. Additional factors affecting duration across languages 18
2.4. NLP models of duration 20
2.5. Interaction effects and the failure of linear models 22
2.5.1. Prosodic position x accent 25
2.5.2. Prosodic position x lexical stress 25
2.5.3. Lexical stress x accent x vowel height / coda voicing 26
2.5.4. Accent x vowel height x coda voicing 27
2.5.5. Prosodic position x coda voicing 28
2.5.6. Vowel tenseness x coda voicing 29
2.5.7. Coda voicing x coda manner 29
2.5.8. Phonemic length x syllable structure 29
2.6. An empirical generalization: hyperadditive lengthening 29
2.7. The task at hand 32
3. Phonetic constraint grammars 34

3.1. Proposing phonetic constraints 35
3.2. Classical OT with categorical constraints 36
3.3. Harmonic grammar with gradiently violable constraints 38
3.3.1. Flemming's parabolic constraints 39
3.3.2. A note on GEN and EVAL 43
3.3.3. Parabolic constraints on duration and timing 45
3.3.4. Targetless constraints with heterogeneous violation functions 54
3.4. A taxonomy 60
4. Maximum entropy phonetic harmonic grammars 66
4.1. Using maxent for phonetics 67
4.1.1. Computing probabilities in maxent phonetic grammars 68
4.1.2. Computing probability with a discretized candidate set 69
4.1.3. Computing probability density with a continuous candidate set 72
4.1.4. Excursus on the feasibility of continuous candidate sets 76
4.1.5. Constraint violations and phonetic distributions 78
4.2. STRETCH and SQUEEZE: hemiparabolic constraints 82
4.3. The unbounded duration pathology 86
4.4. Illustration of a maxent grammar for duration 88
4.4.1. Constraint synergy 94
4.4.2. Asymmetrical constraints and kurtosis 96
4.5. Summary of findings 100
5. Data 103
5.1. Overview of the experiment 104
5.2. Participants 105
5.3. Methods and materials 105
5.3.1. Stimuli 105
5.3.2. Distractor items 109
5.3.3. Equipment 110
5.3.4. Procedure 111
5.3.5. Forced alignment 111
5.3.6. Annotation and exclusion 112
5.3.7. Statistical tests 114
5.4. Results 120

5.4.1. Category means 121
5.4.2. Linear and log-linear models 124
5.4.3. Interaction effects 125
5.4.4. Excursus on duration vs. log-duration as the dependent variable 129
5.4.5. The shapes of durational distributions 130
5.4.6. Testing the skewness hypothesis 132
5.4.7. Testing the uniform variation hypothesis 132
5.5. Discussion and summary of findings 135
5.5.1. The Hyperadditive Lengthening Generalization revisited 137
6. Maxent phonetic learning 144
6.1. Learning in the phonetic domain 144
6.2. The learning algorithm 147
6.2.1. Constraint definitions 147
6.2.2. Constraints 148
6.2.3. Tableaux construction 150
6.2.4. The objective function 152
6.2.5. Parameter learning 155
6.3. Criteria for assessing models 156
6.4. Weight learning with pre-selected targets 156
6.5. Simultaneous weight and target learning 160
6.5.1. DURATION grammar fragments 161
6.5.2. STRETCH and SQUEEZE grammar fragments 167
6.6. Predictions of learned models 176
6.6.1. Predictions of sample means 176
6.6.2. Predictions of standard deviations and kurtoses 178
6.6.3. Conditioned and unconditioned variation 179
6.7. Discussion 181
6.7.1. Target learning: a post-mortem 182
6.7.2. Variation between learned grammars 186
7. Conclusions 187
7.1. The long and the short of it 187
7.1.1. Hyperadditive lengthening 187
7.2. Maxent phonetics 188

7.2.1. The Consistent Variation Hypothesis 188
7.2.2. Constraint synergy 189
7.2.3. Phonetic constraints and phonetic distributions 190
7.2.4. Phonetic learning 190
7.3. New research directions 191
7.3.1. Empirical 191
7.3.2. Theoretical 192
7.4. Outlook 193
Appendix: experimental results 194
References 198

LIST OF FIGURES

Figure 1: Table III from Klatt, 1976, p. 1217. 21
Figure 2: Interaction Results (Klatt 1973b, p. 1103). 24
Figure 3: Figure 3 from de Jong, 2004 (p. 503). Average vowel durations for vowels before voiced and voiceless stops (left panel) and for /æ/ and /e/ (right panel). Plotted here are interactions between focus condition (x-axes) and stress (symbol size). Error bars indicate standard errors. 27
Figure 4: Figure 1 from Choi et al. (2016), p. 624. Effects of coda voicing on vowel duration. (A) Voicing × Focus interactions; (B) Voicing × Focus × Vowel type interactions, as produced by (1) native speakers of English, (2) Korean advanced learners of English, and (3) Korean intermediate learners of English (∗∗∗$p < 0.001$, ∗∗$p < 0.01$, ∗$p < 0.05$). 28
Figure 5: Figure 2 from Boersma 2009, p. 60. 36
Figure 6: Boersma 2009, p. 30 37
Figure 7: The sum of two constraints' violation functions. Red: cost of violating C1(x) where T1 = 5, w1 = 4. Blue: cost of violating C2(x) where (T2 = 15, w2 = 1). Orange: the total cost C(x). Note that the minimum of the total cost function does not lie at the target of either constraint, but in between the two targets. 40
Figure 8: Figure 3 from Flemming, 2001, p. 21. Cost plotted against F2(C) and F2(V). The minimum is located at F2(V) = 1233 Hz, F2(C) = 1467 Hz. 43
Figure 9: Figure 3.4. from Katz, 2010 (p. 112). Data from the production experiment (left) and model predictions (right) for consonant manners with high (rightmost bars) and low (center bars) vowel-recoverability coefficients. For production data, durations are in seconds. The upper bars for vowel-initial items represent closure and transition durations, in realizations where these categories are applicable. 49
Figure 10: (8) from Braver, 2013 (p. 127). Three degrees of phonetic vowel length in Japanese. 50

Figure 11: Figure 12 from Flemming & Cho 2017 (p. 20): “Schematic illustration of the conflict between realizing the magnitude, slope, and alignment targets for a rising tone. The dashed lines show the shape of the rise that satisfies the targets for rise magnitude and slope, while the solid lines schematize the actual slope and magnitude of the rise appropriate for the illustrated intervals between the alignment targets A_L and A_H.” 53

Figure 12: Figure 2 from Windman et al., 2015 (p. 82). Model architecture. Cost functions D (utterance level), P_W (word prominence; only shown for accented word W_{ACC}, as parameter x_{wj} is set to 0 elsewhere) and E/P_S (syllable 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 SUUSUUUS sequence. The y-axes show the costs as a function of duration (x-axis). 56

Figure 13: Figure 8 from Windmann et al., 2015 (p. 83): "Solid lines: cost function *C* (excluding P_W) for stressed (black) and unstressed (gray) syllable with above parameter settings, with circles marking optimal durations. Dashed lines: partial cost functions P_S for stressed (black) and unstressed (gray) syllable with above parameter settings.” 57

Figure 14: Figure 7 from Windmann et al., 2015 (p 85): “Absolute (left) and proportional (right) amount of accentual lengthening in stressed and unstressed syllables in the simulated utterance (bisyllabic accented word). 58

Figure 15: Taken from Hayes & Schuh (MS), p. 38; adapted from Flemming & Cho (2017): Violation functions for two constraints and their summed violations, comparing parabolic with linear violation functions. Linear violation functions fail to predict compromise between targets. 64

Figure 16: Violations, maxent values, and probabilities as a function of duration for a single-constraint maxent grammar with a discretized candidate set 72

Figure 17: Violations and maxent values as a function for duration for a single-constraint maxent grammar with a continuous candidate set. 73

Figure 18: Predicted distribution for a single-constraint maxent grammar with a continuous candidate set. 75

Figure 19: Maxent value as a function of F2(C) and F2(V) (*x* and *y*, respectively) in a maxent grammar with ID(V), ID(C), and MINEFFORT, with the consonant F2 locus set to 1700 Hz, the vowel F2 target set to 1000 Hz, and all weights set to 1, following Flemming (2001). Durations are represented in hectohertz (10^2 Hz), rather than hertz, for visual clarity. 77

Figure 20: Harmony and maxent value functions in a hypothetical maxent grammar with the constraints proposed by Windmann et al. (2015). 80

Figure 21: Hemiparabolic violation functions for STRETCH (left) and SQUEEZE (right), with weights set to 1 targets set to 7.5. 84

Figure 22: Violations incurred by a STRETCH constraint with a high weight (red) and a SQUEEZE constraint with a low weight (blue), and the maxent values for candidates subject to both constraints (green). The resulting distribution has a positive kurtosis / skew. 85

Figure 23: A “soft” window model of duration, composed of a STRETCH constraint with a shorter target (red) and a SQUEEZE constraint with a longer target (blue). Maxent values for candidates subject to both constraints are shown in green: the predicted distribution (at least

without any other constraints) is uniform free variation within a duration range, with a small number of outliers, depending on the constraint weights. 86

Figure 24: A pathological grammar with only a STRETCH constraint. Since the area under the maxent value curve is not defined, the probability distribution cannot be calculated. 87

Figure 25: maxent duration à la Klatt. Left: the violation functions for STRETCH[V] + SQUEEZE[V] (black), STRETCH[V/__[-voice]] (blue), and SQUEEZE[V/__σ] (red). right: the probability distributions for [1σ +v] (black), [1σ -v] (blue), [2σ +v] (red), and [2σ -v] (magenta). 93

Figure 26: Probability distributions for the durations of vowels in the contexts [1σ +v] (black), [1σ -v] (blue), [2σ +v] (red), and [2σ -v] (magenta) in two maxent grammars, one where the two category-specific constraints are both SQUEEZE constraints and the longest category is the base case (top), and one where they are both STRETCH constraints and the shortest case is the base case (bottom). 96

Figure 27: Violation profiles for DURATION (blue) and SQUEEZE (red), and predicted duration distributions (green) for grammars with the constraints DURATION and SQUEEZE. All weights are set to 1, the target for DURATION is set to 3, and the target for SQUEEZE is either 3 (top) or 1 (bottom), also shown as a verticle line in red. 99

Figure 28: Means by vowel for just the closed syllable data. Error bars represent 95% confidence intervals. 121

Figure 29: Means by coda for just the tense vowel data. Error bars represent 95% confidence intervals. 122

Figure 30: Means by onsets for all the data excluding targets “bed” and “bets.” Error bars represent 95% confidence intervals. 123

Figure 31: Means by prosodic condition for all the data. Error bars represent 95% confidence intervals. 123

Figure 32: Residuals as a function of predicted duration for the linear (left) and log-linear (right) mixed effects models with only main effects, with loess lines. 128

Figure 33: Histograms and probability density plots for tokens of phrase-medial unaccented /ɛ/ (top) and phrase-final accented /eɪ/ (bottom). Red lines indicate mean durations. 131

Figure 34: Skewness of samples as a function of their mean duration (left) or mean log duration (right) across the 128 experimental conditions. 132

Figure 35: Differences between longer and shorter categories, across each of the nine binary phonological features, in their propensity to show conditioned and unconditioned variation, for both duration (left) and log-duration (right). 133

Figure 36: Two ways of understanding the Hyperadditive Lengthening Generalization. Blue bars indicate durations that would occur if the two effects combined multiplicatively with no interactions, where the longest duration (left) or the shortest duration (right) is predicted on the basis of the other three. Red bars indicate the deviation from this prediction that occurs for many pairs of durational factors, interpreted as unexpectedly long durations for either the category undergoing two lengthening effects (left) or two shortening effects (right). 139

Figure 37: The means of the duration distributions predicted by the S_{sparse} grammar compared with observed sample means from the 128 experimental conditions. 176

Figure 38: Predicted distributions of the S_{sparse} grammar (red), and observed histograms from the training data (blue), for six of the 128 experimental conditions. 177

Figure 39: The standard deviations of the duration distributions predicted by the S_{sparse} grammar compared with observed sample standard deviations from the 128 experimental conditions. 178

Figure 40: The standard deviations of the kurtoses predicted by the S_{sparse} grammar compared with observed sample kurtoses from the 128 experimental conditions. 179

Figure 41: Predicted differences between longer and shorter categories, across each of the nine binary phonological features, in their propensity to show conditioned and unconditioned variation, for both duration (left) and log-duration (right), according to the S_{sparse} grammar. 180

LIST OF TABLES

Table 1: Mean English vowel duration by accentuation and phrase-finality reported by Li & Post (2014). 25

Table 2: Reported "interactions" between effects between factors affecting vowel duration. All interactions were in the positive direction (if the effects are treated as both being lengthening or both being shortening effects), except for the interaction between pitch accent and phrasal position, and potentially the interaction between coda manner and coda voicing. 31

Table 3: (11) from Flemming, 2001, p. 19 : Cost functions for the three constraints. 42

Table 4: Table I from Flemming, 2001, p. 21. Evaluation of example candidate values for F2(C) and F2(V), with L = 1700 Hz, T = 1000 Hz, and all weights set to 1. 42

Table 5: Constraints and cost functions used by Windmann et al. (2015), where… s_i is the duration of the *ith* syllable, w_j is the duration of the *ith* word, η_i is an effort coefficient for the *ith* syllable, ψ_i and ψ_j are perceptibility coefficients for the *ith* syllable and *jth* word, α_{wj} is a coefficient for the strength of word prominence in the *jth* word, and δ_i is a speech rate coefficient for the *ith* syllable. 54

Table 6: A typology of some possible formulations of phonetic constraint grammars. Where cells are divided into pairs, the top one represents an implementation with a discretized candidate space. Shaded areas denote theoretical combinations judged impossible. 62

Table 7: Maxent grammar example. 68

Table 8: Violations maxent values, and probabilities for a single-grammar constraint with a discretized candidate set. 71

Table 9: Definitions of the durational constraints STRETCH and SQUEEZE. 83

Table 10: Observed and predicted vowel durations, by context, in Klatt, 1973b. 89

Table 11: Constraints for a grammar governing the interaction between two durational effects. 90

Table 12: Predicted mean vowel duration, using the maxent grammar. 93

Table 13: Grammars with two shortening constraints (left) and two lengthening constraints (right) based on the experimental results from Klatt 1973(b)..... 95

Table 14: Target words..... 106

Table 15: Proper names used in carrier sentences. 108

Table 16: The four prosodic frames, using “bed” as the target word, and “Susan” as the proper name..... 109

Table 17: Examples of distractor items. 110

Table 18: Binary features used as fixed effects in the linear and log-linear regressions..... 114

Table 19: For each of the binary features in the experiment, the pair-wise matched sets of experimental conditions which were compared to test the consistent variation hypothesis.117

Table 20: Means and standard deviations of the durations in the corresponding pre-voiceless and pre-voiced experimental conditions. Pre-voiceless conditions with no corresponding pre-voiced condition, namely those involving the targets “met,” were excluded from the comparison..... 120

Table 21: The fitted parameters of linear and log-linear mixed effects regressions on the data, with a random variable of speaker..... 124

Table 22: Differences between longer and shorter categories, across a number of phonological features, in their propensity toward conditioned variation, and toward unconditioned variation, for both duration (left) and log-duration (right). 133

Table 23: Reported “interactions” between effects between factors affecting vowel duration. A (*) indicates that a significant interaction ($|t| > 2$) was found, and a (.) indicates that a trend ($|t| > 1$) was found, in the log-linear models of the experimental data from this chapter. All interactions were in the positive direction (if the effects are treated as both being lengthening or both being shortening effects), except for the interaction between pitch accent and phrasal position. 138

Table 24: A superset of the constraints used in any one learning attempt..... 150

Table 25: Hypothetical example of a tableau prepared during initial tableaux construction, prior to learning, with duration range 0 - 500 ms and duration resolution 20 ms. 152

Table 26: Hypothetical example of a maxent tableau created during learning..... 154

Table 27: Learned weights for constraint set D_{full} with targets pre-set to the means of the natural class they constrain. Weights of 0 are omitted for visual clarity..... 158

Table 28: Learned weights for constraint set D_{full} with targets pre-set to the mean (black), the mean plus 1 SD (red), or the mean minus 1 SD (blue) of the natural class they constrain. Weights of 0 are omitted for visual clarity..... 159

Table 29: Learned weights for constraint set D_{full} with targets pre-set to the mean (black), the mean plus 2 SDs (red), or the mean minus 2 SDs (blue) of the natural class they constrain. Weights of 0 are omitted for visual clarity. 160

Table 30: Learned weights and targets (in deciseconds) from four training runs for constraint set D_{full}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted—0.00 indicates a positive value smaller than 0.005. 162

Table 31: Learned weights and targets (in deciseconds) from training run #2 of constraint set D_{full}, with mean durations of the class constrained for comparison. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). 0.00 indicates a positive value smaller than 0.005. 163

Table 32: Learned weights and targets (in deciseconds) from four training runs on constraint set D_{short}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). 165

Table 33: Learned weights and targets (in deciseconds) from training run #1 on the constraint set D_{short}, with mean durations of the class constrained for comparison. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). 166

Table 34: Learned weights and targets (in deciseconds) from four training runs, using constraint set D_{long}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted—0.00 indicates a positive value smaller than 0.005. 167

Table 35 : Learned weights and targets (in deciseconds) from four training runs, using constraint set S_{full}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted—0.00 indicates a positive value smaller than 0.005. 169

Table 36: Learned weights and targets (in deciseconds) from four training runs, using constraint set S_{sparse}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted—0.00 indicates a positive value smaller than 0.005. 171

Table 37: Learned weights and targets (in deciseconds) from training run #2 on the constraint set S_{sparse}, with mean durations of the class constrained for comparison. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted. 172

Table 38: Learned weights and targets (in deciseconds) from four training runs, using constraint set S_{short}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted—0.00 indicates a positive value smaller than 0.005. 174

Table 39: Learned weights and targets (in deciseconds) from four training runs, using constraint set S_{long}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted—0.00 indicates a positive value smaller than 0.005. 175

Table 40: Model performance for all the constraint sets learned..... 175

Table 41 : Predicted differences between longer and shorter categories, across a number of phonological features, in their propensity toward conditioned variation, and unconditioned variation, for both duration (left) and log-duration (right), according to the S_{sparse} grammar. 180

ACKNOWLEDGEMENTS

First and foremost, I am indebted to Bruce Hayes, without whom I would never have thought to apply the tools of maxent OT to the problem of duration, or indeed known very much about phonological theory at all. At every stage of producing this dissertation, he helped me stay on track, think clearly, and keep the bigger picture in mind when I got sidetracked or distressed by some detail or other. Beyond academic mentorship, I've been a long-time beneficiary of Bruce's deep concern for the well-being of his students and his department: I am grateful to him for providing me with research assistantships, teaching positions, recommendation letters, bureaucratic help, encouragement, and life guidance whenever I needed these things.

Many thanks are owed to my dissertation committee for their advice, their support, and most of all for their patience—for a dissertation about time constraints, this project took a rather long time to complete.[1] Even before its inception, though, these individuals helped sow its seeds by providing me with a world-class linguistics education.

Sun-Ah Jun introduced me for the first time to the wide world of intonation and prosody, and advised me in the design of most of the experiments I conducted during my graduate school career, saving me from numerous potential pitfalls.

Robert Daland's pro-seminar on maxent grammars helped me to understand math that would be turn out to invaluable to this dissertation, and simultaneously gave me valuable research programming experience, which serves me to this day.

Bruce Hayes and Kie Zuraw gave me the best phonology education that any student of linguistics could ask for.

---

[1] About 132,710,400,000 ms. But who's counting?

Many other members of UCLA linguistics department have been at one time or another a sounding board for many of the ideas presented in this dissertation, and the source of insight or suggestions that helped shape it into its current form. I am grateful to them for this, but even more so for being such a warm, friendly, intelligent group of people, who I have been honored to call my friends, colleagues, and teachers for so many years.

Finally, this dissertation wouldn't have been possible without my parents, Larry Lefkowitz and Judy Snape. This is true for a rather obvious biological reason. Beyond that, it is true because they instilled in me the love for learning, intellectualism, and self-confidence without which I would never have made it this far.

VITA

2010 B.A. in Linguistics, *Cum Laude*

Dartmouth College

Hanover, NH

2013 M.A. in Linguistics

UCLA

Los Angeles, CA

1. Introduction

1.1. The case for phonetic grammars

This dissertation uses phonetic duration as a case study in developing a particular type of formalism for the portion of the grammar which governs speakers' knowledge of phonetics, as distinct from phonology. The need for such a formalism is not historically uncontroversial. In their seminal work, Chomsky and Halle (1968) justify the use of phonetic symbols to the exclusion of phonetic measurements in their grammars by positing that the details of phonetic implementation are subsumed by general physiological facts about motor planning, are not language specific, and are therefore extra-grammatical:

> Even if the phonetic transcription were as faithful a record of speech as one could desire, there is still some question whether such a record would be of much interest to linguists, who are primarily concerned with the structure of language rather than the acoustics and physiology of speech. It is because of this that many structural linguists have felt that phonetics has very little to offer them and have therefore assigned to it a secondary, peripheral role. (p. 293)

> ...in [our] view, phonetics is concerned with grammatically determined aspects of the signal....thus it is no longer a problem that the transcription is composed of discrete symbols whereas the signal is quasi-continuous, or that the transcription provides information only about some properties of the signal and not about others. (p. 294)

Even if we concede that linguists should only be concerned with grammatically determined aspects of the signal, Chomsky and Halle make the crucial assumption that grammars only ever need manipulate discrete symbols, and that, in the process of producing from these symbols the

quasi-continuous signal that is speech, only general, physiological, extra-grammatical processes are involved.

As others have argued (Keating, 1985; Flemming, 2001), these assumptions turned out to be unfounded: non-categorical patterns of phonetic targets and phonetic variation, such as the degree and timing of final devoicing (Keating, 1985), differences between the inherent duration of vowels (Lehiste, 1970; Westbury and Keating, 1980; Maddieson, 2004), pre-voiced-obstruent vowel lengthening (Keating, 1979; Klatt, 1973; Lehiste, 1970), and the degree of overlap or articulation between adjacent consonants (Zsiga, 2000), to name just a few,[2] occur to different degrees in different languages. Because languages' sound systems can differ by phonetic degrees, and not just categorically, some continuous phonetic parameters must be language-specific, and therefore learned. Unless these patterns reside solely in the lexicon, they must be governed by some part of the grammar.

Exactly how this phonetic component of the grammar should be studied or represented is an as yet unanswered (and, too often, unasked) question. A few explicit proposals do exist: in the domain of intonation and phrase-level prosody, Pierrehumbert (1980; 1981) and Beckman and Pierrehumbert (1986) use phonetic rules to transform ToBI-style prosodic transcriptions into a discrete number of real-number value articulatory targets, which are then connected by an interpolation mechanism, and may be under- or overshot. Keating (1990a) proposes a more flexible model in which, in any particular articulatory dimension (e.g. velum height), sounds or phonetic categories have grammatically determined ranges, or "windows", rather than targets, within which

[2] It is interesting that these examples, some of the most clear-cut arguments for language-specific phonetics, are all related to duration and timing in one way or another.

variation is permitted, and this approach is adopted by later authors in work on segmental alignment (Byrd, 1996b; Zsiga, 2000). Alternatively, the approach in the Articulatory Phonology framework (Browman & Goldstein, 1992) is to do away with much of the phonological component as traditionally imagined, replacing phonological features with articulatory gestures. Both phonetic and phonological processes in this model apply to temporally extended muscular events, which have continuous representations from the outset. Regardless, a holistic, formal, predictive, and explanatorily adequate account of how phonetic targets are mathematically determined by the grammar is needed.

In this dissertation, I will investigate the line of reasoning that *maximum entropy harmonic constraint grammars*, already in use by phonologists, are suitable for modeling phonetic knowledge and phonetic variation.

1.2. Duration as a testbed

Phonetic duration, defined as the size in the time dimension of individual speech events (be they segments, gestures, or larger prosodic constituents), is a fertile testbed for research on the nature of phonetic grammars for several reasons. First, it is comparatively easy to measure: all segments and larger prosodic constituents have durations, and some segment boundaries are reliably observable from acoustic data alone. Second, duration uncontroversially exists in both the articulatory and the acoustic domain, unlike other acoustic or articulatory phonetic dimensions. This fact allows me to remain agnostic on the highly controversial question of whether production grammars involve phonetic representations that are fundamentally articulatory or acoustic. In other words: time is time.

Last, and perhaps most importantly, many phonological factors are known to affect duration (such as segmental context, accentuation, etc.), and the effects of each of the factors individually (in English and few other languages) are already well-documented (see Chapter 2 of this dissertation for an overview). However, the way in which these factors interact is not well understood, and capturing any such interactions is exactly the job of a phonetic grammar. In this dissertation, the eventual durations of segments or prosodic constituency subject to multiple, sometimes opposed, duration-related effects will be modeled as the result of constraint interaction.

1.3. Roadmap of the dissertation

Chapter 2 defines phonetic duration, and gives a literature review of the many segmental, prosodic, intonational, and extra-grammatical factors that are already known to influence duration, and also reviews a few holistic mathematical models of duration that have been used by speech synthesis researchers. A surprising generalization regarding how various factors can be seen to interact in determining duration is found.

Chapter 3 reviews the small literature on phonetic constraints and phonetic constraint grammars, providing a typology of the possible approaches to adapting phonological constraint grammars to the domain of phonetics.

Chapter 4 explores how maxent grammars could be used to model the realization of continuous phonetic variables like duration, discusses potential constraints, violation functions, spaces of candidates, and evaluation functions that these grammars might employ, demonstrates using toy examples how such a grammar would work in practice, and argues that the proposed framework and constraints in fact make a number of interesting and testable empirical predictions regarding the behavior of the phonetic variables they govern.

Chapter 5 presents the results of a production experiment that investigates the nature of the interaction between various phonological factors known to affect duration. The experiment investigates the duration of English front vowels in monosyllables with varying segmental, prosodic, and intonational contexts. The experimental results are shown to confirm the generalization from Chapter 2, and, encouragingly, to be consistent with several theoretical predictions from Chapter 3. The results are also used as training data for a maxent learner developed in Chapter 6.

Chapter 6 adapts maxent learning algorithms of the sort already used by phonologists to the phonetic domain, discussing certain theoretical and implementational challenges that present themselves, and applies this adapted learning algorithm to the experimental data from Chapter 5. Investigation of learned grammar fragments reveals that choices as to which model parameters are set in advance, and especially which constraints and constraint families are used, affect the ability of the resulting grammars to successfully model the data. Grammars consisting of constraints which involve phonetic targets for duration are found to fit the data well, so long as the learner is allowed to manipulate these targets. However the target values it learns are often counterintuitive, and shed new light on how phonetic constraints and constraint parameters should be interpreted.

Chapter 7 summarizes the main findings of the dissertation, and suggests directions for future research.

2. Duration as a phonetic variable

2.1. Duration defined

Duration as defined here refers to the size in the time dimension of speech events. In the empirical work that is to follow, I will choose to focus on events that correspond to segments, but focus only on segments that are, in the right context, well-defined acoustically—in other words, speech segments that are internally homogeneous in some way and have relatively clear acoustic boundaries that separate them from the immediately preceding and following material.

That the duration in time of these sorts of acoustic events in particular is in fact relevant for speakers and listeners is intuitive. Sounds are considered by most to be the starting point to perception and recognition, and since we presumably "speak in order to be heard, and need to be heard in order to be understood" (Jakobson & Waugh, 1979), it would be advantageous for speakers to have a fairly good idea of what noises they are trying to make. Explicit arguments for the centrality of acoustic representations to both perception and production are not hard to find (e.g. Keating, 1990b; Boersma, 1998), and speakers can even make drastic, ad-hoc adjustments to their articulatory strategies to meet acoustic / auditory goals when the articulators are physically impeded (Mayer et al., 2009).

Nevertheless, this is a matter of some contention. Linguists who take the basic units of speech to be articulatory gestures, rather than featurally defined phones, consider the relative timing of individual gestures and the amount of overlap in time of such gestures to be central, and the particulars of the acoustic correlates of these gestures to be largely irrelevant for production, and important in perception only in so far as they aid in the recovery of articulatory gestures (Browman & Goldstein, 1992; Gafos, 2002). A few espouse an "ecological" or "direct realist" approach to

perception (Best, 1995), taking a more extreme view: that acoustic information is never grammatically represented and is not a part of language cognition at all, but that listeners instead have direct access to the articulatory events themselves.

I will not weigh in on this ongoing debate in this dissertation. To those in the articulatory camp who see the decision to measure acoustic events as a fatal flaw, dooming my empirical work and subsequent model-fitting to failure and uninterpretability, I justify my decision as follows: in some cases, articulatory events and acoustic events are fairly well aligned (for example, stop closures and releases), allowing the times of acoustic measurements to serve as accurate surrogates for the times of certain corresponding articulatory events. Of course, whether articulatory events such as points of contact or release for stops are the "right" events to be studying is itself a point of contention, in part due to issues like articulatory under- or overshoot. Furthermore, the duration of an articulatory event thus defined is not necessarily a primitive in the grammar: for example, a jaw-lowering gesture will be longer if it is performed more slowly, but will also be longer if the target is a lower jaw height, or if it is performed with less gestural stiffness, and in fact these articulatory parameters are known to be somewhat independently targetable by different prosodic factors (Edwards et al., 1991; Byrd & Saltzman, 1998).

Nevertheless, consistent, phonologically-determined *patterns* of phonetic duration in the acoustic signal should reflect qualitatively similar patterns of variation in the timing or size of articulatory gestures, so any qualitative patterns discovered with respect to significant duration differences between categories which are phonologically minimally different should be theory-neutral empirical results.

A more substantive problem is the following: it is unclear to what extent the acoustic or articulatory representations in the minds of speakers and listeners "match" their actual productions

and the acoustic signals picked up by a microphone, since physical articulatory and auditory systems intervene between these grammatical representations and the observable performance data. Once again, for convenience, I will make the simplifying assumption that speakers' phonetic representations of duration generally match their performance, at least qualitatively: when a sound in some environment or with some property is systematically longer or shorter, and especially when the presence or extent of this effect is known to be language specific, this variation should be explainable with reference to some part of the grammar. Formal work in physiology and in language processing could alleviate the need for this simplifying assumption by combining explicit physical models of the mechanical, extra-grammatical components of the speech apparatus (see, for example, Vogt et al. 2005) with the models of acoustic duration targets presented here, but this is beyond the scope of this dissertation.

Finally, it's not clear on what scale or with which units duration should be represented in the grammar, or whether these representations have a linear relationship with time in the actual speech signal (Katz, 2012). Many perceptual scales in language (loudness, pitch, vowel formants) relate to the corresponding physical dimension (intensity, f0, formants) in a roughly logarithmic way (e.g. the Mel scale for pitch; Stevens & Volkmann, 1940), and it is quite possible that duration may behave this way as well. With respect to perception, for example, Small and Campbell (1962) find that the just noticeable difference (JND) for non-linguistic stimuli (silence, a tone, or noise) increases logarithmically with the duration of the stimuli, meaning listeners are more attuned to durational differences when comparing stimuli of shorter duration. A linguistic argument for modeling duration logarithmically is provided by Rosen (2005), who argues that segment durations tend to obey lognormal distributions, and that effects on duration are proportional (log-linear)

rather than additive. In the empirical and model-fitting chapters of this dissertation, therefore, both duration and log duration will be considered potential outputs of the grammar.

2.2. Factors affecting duration in English

In this section, I give a brief literature review of the many linguistic factors that might affect the duration of English speech sounds, including the segmental features of the sounds themselves, lexical prosody, sentence-level prosody, lexical features, discourse factors, and speech rate. Additionally, I review what is known about how some of these factors interact.

2.2.1. Segmental factors

2.2.1.1. Segmental features

Even in languages which are thought to lack a phonological length contrast, such as English, different segments and their subparts can have systematically different durations. For example, English fricatives are much longer than stops (Klatt 1973a), at least in onset position, and the lax vowels of English are significantly shorter than the tense vowels, with further differentiation by vowel height (Peterson & Lehiste, 1960). Klatt (1975) finds that among the stressed vowels of English, as much as half of the total variance in duration in fluent speech can be attributed to segment identity alone. Small but reliable phonetic differences in VOT between stops at different places are also widely attested, and even argued to be universal (Cho & Ladefoged, 1999).

A popular way to account for phoneme-by-phoneme durational variation is to posit that each phoneme (or perhaps each allophone, or each gesture) of a particular language has an "intrinsic" or "inherent" duration (Lehiste, 1970, 1975a). This is the approach taken by essentially everyone interested in fully predictive models of duration (Klatt, 1973b, 1976; en, 1997; van Santen et al.,

1997). However, reporting the mean duration of each phoneme, while perhaps an observationally adequate approach, misses various generalizations that can be made about subsets of phonemes. For example, it seems likely that speakers know that vowels are generally longer than consonants, that "tense" vowels are longer than "lax" vowels, that low vowels are longer than high ones, and so forth. In the "intrinsic duration" account, in so far as target durations for each phoneme are simply stored in a list, these generalizations are treated as being incidental, or at least are not encoded anywhere in the model. This is a drawback that needs to be addressed if our goal is a descriptively adequate, human-like grammar, and not merely a speech synthesis system. Even from an empirical standpoint, it seems possible that, upon encountering a foreign word with an unfamiliar phoneme, the duration (and other phonetic properties) of that borrowed phoneme could be influenced by its membership in natural classes in the native language, just as its phonological behavior can be so influenced (Halle, 1978).

2.2.1.2. Segmental context

The features of neighboring segments can influence segmental duration. The voicing of a following obstruent affects the durations of vowels (e.g. Lehiste, 1970; Klatt 1973b, 1976; Crystal & House, 1988; De Jong, 2004), so much so that vowel duration is a strong acoustic cue for coda voicing (e.g. Crowler & Mann, 1992; Moreton, 2004). The manner has an effect as well, with coda nasals associated with longer vowels than coda stops (Umeda, 1975; Katz, 2010; Crystal & House, 1988),[3] as does its place of articulation, with longer vowels occurring before bilabial codas than before alveolar or velar ones (Luce & Charles-Luce, 1985; Crystal & House, 1988).

[3] The duration of vowels before stops as compared to fricatives has conflicting results in the literature, with longer vowels before stops reported by several authors, but with subsequent failures to replicate this finding, and

The features of preceding onset consonants also affect vowel duration, with shorter vowels appearing after voiceless onsets than after voiced ones, and after sonorant onsets than after obstruent ones (Crystal & House, 1988; Katz, 2010).

2.2.2. Word-internal prosodic factors

2.2.2.1. Syllable structure / compensatory shortening

The complexity of a syllable and of its constituent parts, as well as the phonological properties of the segments within a syllable, can have a large effect on the duration of its segments. The most well-known example is closed syllable vowel shortening: vowels in many languages differ in length depending on whether they are in an open or a closed syllable, even controlling for segmental environment (e.g. in pairs like “beak age” and “bee cage”).

Katz (2010, 2012) thoroughly investigates compensatory vowel shortening in English, and finds that both onsets and codas induce vowel compression, and furthermore that adding additional consonants to create branching onsets or codas sometimes but not always induces an additional compensatory shortening effect on the vowel, termed “incremental compression,” depending on the quality of the consonant closest to the vowel. In particular, he finds that:

> All consonants are associated with some amount of simple vowel-compression, but not all strings induce incremental compression. Clusters including liquids induce incremental compression in both onset and coda position relative to liquid singletons, clusters including nasals do so only in onset position, and clusters containing only obstruents do not condition

with large differences depending on the phrasal-position in which these comparisons are made, among other factors (Chrystal & House, 1988).

incremental compression in either position. For instance, the vowels in /brod/ and /dɔrb/ are significantly shorter than those in /rod/ and /dɔr/, but the vowel in /donz/ is not shorter than that in /don/. (Katz, 2012; p. 7).

Katz posits that these facts are the direct result of a perceptual asymmetry: consonants like liquids, which exhibit longer and more audible coarticulation with the vowel, contain in them cues for the vowel, and so the vowel proper can be shortened without much harming recoverability. Asymmetries between onset and coda position on incremental vowel compression are similarly explained.

Consonants themselves have different durations depending on whether they appear in onset or coda position, and can also exhibit compensatory shortening, decreasing in duration as the syllable onset or coda becomes more "crowded," with some exceptions. In onsets, for example, increasing the number of consonants generally shortens all of the onset consonants (compared to their duration in singleton onsets), but those closest to the nucleus are shortened the most (Klatt 1973a, 1974). Additionally, compensatory shortening affects different consonants differently, with labial consonants, for example, being relatively incompressible, forcing consonants which share a cluster with a labial (like the [l] in 'kelp' or the [ɹ] in 'pry') to shorten more than they would in a cluster without a label in order to accommodate this incompressibility (Klatt 1973a). Byrd and Tan (1996) demonstrate that, articulatorily speaking, compensatory shortening of consonants in clusters is achieved via a combination of both shortened consonant gestures and increased overlap between consonantal gestures (when such overlap is possible), a result that underscores the significance of the "duration vs. timing" distinction.

These compensatory shortening results very strongly suggest that prosodic constituents larger than segments, such as syllables, have target durations, a hypothesis that will be explored in the chapters that follow.

2.2.2.2. Lexical stress

In English, the vowels of syllables with primary stress[4] are longer than those in syllables with secondary stress, which in turn are longer than unstressed vowels (e.g. Klatt, 1976; De Jong, 2004). Additionally, consonants are longer in the onsets of stressed syllables than they are in the onsets of unstressed syllables (Oller, 1973). English consonants' durations can also show sensitivity to stress. For example, English /s/ in onset position has a shorter and more variable duration in stressless syllables than in stressed syllables, where it is both longer and more stable (Klatt, 1974).

2.2.2.3. Word length

Stressed syllables in disyllabic trochaic words are shorter than identical syllables in monosyllabic words, at least in English (Klatt, 1973b). In fact, this shortening effect applies not only to the vowel, but to (at least) onset consonants as well (Klatt, 1973a, 1974).[5] Once again, this compression suggests that larger prosodic constituents, such as the prosodic word, have target durations.

---

[4] It is worth mentioning here that the location of English stress is not always entirely lexical, as evidenced by metrical phenomena like the "rhythm rule" of English (cf. 'fifteen men' vs. 'nine-fifteen').

[5] This result could be attributed to word-final lengthening, rather than shortening in polysyllabic words, so long as the domain of word-final lengthening includes the onset consonants of the final syllable. These possibilities could be distinguished empirically by running the same experiment with iambic words included.

2.2.3. Phrase-level prosodic factors

2.2.3.1. Prosodic location

The right edges of a various prosodic domains, such as a prosodic word, accentual phrase, intermediate phrase, or intonation phrase, are associated with lengthening (Lehiste, 1972; Byrd & Saltzman, 2003), and more lengthening occurs in larger domains. At least four such domains, each associated with a different degree of lengthening, are relevant for English (Wightman et al; 1992), and this lengthening affects the duration of consonants (e.g. Klatt, 1974) as well as vowels.

Conversely, syllable-, word-, phrase-, or utterance-initial positions are associated (in English and perhaps universally) with articulatory strengthening and fortition (Keating et al., 2003). This fortition can, directly or indirectly, result in longer consonant durations in domain-initial positions. For example, Keating et al. find that English /n/ has both more peak articulatory contact and a longer contact duration when it is domain initial, and the degree of lengthening depends on the type of prosodic domain it initiates.

2.2.3.2. Accentedness

Pitch-accented syllables have longer duration than lexically stressed but unaccented syllables (e.g. Anderson et al., 1984). Duration, along with amplitude, are primary cues for accentedness in English (Turk & Sawusch, 1993).

2.2.4. Lexically specific and discourse-level factors

In this section, various factors that affect phonetic reduction or phonetic hyperarticulation in general are discussed, since reduction and hyperarticulation can involve changes in segment duration. It is not always clear whether these factors belong in the grammar *per se*; they may

instead be universal and/or be attributable to facts about language processing. For example, Baese-Berk & Goldrick (2009) posit that lexical neighborhood effects in production, and possibly lexical effects in general, are best explained by way of competition that occurs during lexical access.

It is worth noting that in some models (e.g. Pierrehumbert, 1981), focus and other discourse-level factors contribute to *prominence*, a numerical value assigned to words or phrases. This real-number value, as interpreted by the intonational component of the grammar, in turn affects the pitch ranges of prosodic phrases and the pitch targets of accented words, as well as their durations. Aylett and Turk (2004; 2006) posit that the expected retrievability of a word in context is in fact the single most important factor for durational variance in production: in their estimation, the primary function of intonation is to spread informativity evenly over an utterance, an idea they call the "smooth signal" hypothesis.

2.2.4.1. Focus

Focus in English is marked by a number of intonational factors, including increased prosodic prominence, which is realized in part through lengthening.

2.2.4.2. Discourse salience

Tokens of words that are more salient or more discourse-given are reduced compared to those that are less salient or discourse-new (Hawkins & Warren, 1994; Fowler, 1988). It is unclear, however, to what extent this effect is distinct from that of discourse level focus more generally.

2.2.4.3. Lexical frequency

High-frequency words show more phonetic reduction than phonologically similar low-frequency words (e.g. Pluymaekers *et al.,* 2005). This may or may not be merely a special case of contextual predictability (see below).

2.2.4.4. Lexical neighborhood density

Words for which there are many similar-sounding words or "lexical neighbors", which are as a result more confusable with other words, are produced with less reduction (Wright, 2004), and with more consonant-vowel coarticulation (Scarborough, 2004), both of which could affect segment duration.

2.2.4.5. Contextual predictability

Tokens of words which are predictable from context are reduced compared to tokens which are not predictable (Lieberman 1963; Aylett & Turk 2006).

2.2.4.6. Location in a discourse

Some results from reading-style speech show that the last sentence of a paragraph will be read more slowly than other sentences (Lehiste, 1975b). However, this is almost certainly best described as a global change in speech rate.

2.2.5. Speech rate

While changes in speech rate might be extra-grammatical or even extra-linguistic (Klatt, 1976), faster speech obviously necessitates shorter duration for some or all of the sounds involved.[6]

Perhaps the simplest hypothesis for how this might occur would be that segments shorten proportionally, such that when a sentence uttered at twice the rate, its segments will each be half as long. This is almost certainly not the case. For instance, in slower speech, pauses account for a disproportionate amount of the added time (Goldman-Eisler, 1968), and, conversely, in fast speech prosodic boundaries are not as strongly marked (Fougeron & Jun, 1998), such that phrase-final material is disproportionately affected by changes to speech rate.

In faster speech, consonants and vowels are both shortened, but not to the same degree (Goldman-Eisler, 1968), and finer grained distinctions can be made between sonorants, fricatives, and stops with regard to their duration in faster speech taken as a percentage of their duration in slower speech (Crystal and House, 1988).

In the articulatory domain, Gay et al. (1974) find that, in fast speech, the articulatory gestures for (labial) consonants are actually strengthened, while the gestures for vowels show a decrease in articulatory strength, and are more likely to exhibit articulatory undershoot (Lindblom, 1963). Byrd & Tan (1996) report increased overlap of consonantal gestures in consonant clusters in fast speech, but that the degree to which this overlap happens is dependent on the segmental features of the consonants involved.

---

[6] Unless of course a different spell-out is produced, in which some segments delete entirely or are replaced by free variation allophones which are inherently shorter—a clear interaction between speech rate and phonology. In a gestural framework, increased speech rate could in theory also be effected merely by increasing gestural overlap rather than shortening gestural duration, but empirically there is evidence that both of these strategies are used in conjunction (Byrd & Tan, 1996).

A more plausible hypothesis regarding speaking rate, therefore, is that rate is a "knob" which affects the parameters of some part of the production grammar, changing durational targets in a non-uniform way that depends on the structure of the grammar and the segments and prosodic structures involved. A more specific version of this hypothesis, if the grammar is taken to involve constraints (Chapters 3 and 4), is that the parameters thus affected are constraint weights, or even the weight of a single constraint on the duration of some large prosodic constituent. As far as I know, nobody has directly implemented this hypothesis.

The extent to which different classes of sounds and prosodic constituents are affected differently by changes in speech rate is a largely open question (and one worthy of experimentation!), but it is clear that it cannot be considered independently from the grammar when production data is being modeled.

2.3. Additional factors affecting duration across languages

While a thorough overview of the typology of phonetic duration is far beyond the scope of this dissertation, it is worth mentioning a few phonological features which are absent from English, but which have an effect on duration in the languages where they do occur.

In many languages, length is a contrastive phonological feature. Generally, the distinction is a binary one, between singleton and geminate consonants or vowels, but three-way contrasts in vowel and consonant length are attested (though exceedingly rare), appearing in languages such as Estonian (Prince, 1980), and Dinka (Remijsen & Gilley, 2008).[7] It goes without saying that

[7] Less controversially, three-way phonological contrasts in voice onset time, effectively a subsegmental duration or timing contrast, are widely attested.

phonetic duration is affected by phonological length, but the phonetic realization of phonological length can be language-specific (Smith, 1993), and is therefore arguably part of the phonetic component of the grammar.

In languages like Japanese, where length is indicated orthographically, written words can sometimes be emphasized by elongating sounds in them to an arbitrary degree ("no" vs. "nooo" vs. "noooooooo"). In these cases, greater than ternary length "contrasts" can purportedly be produced by speakers and distinguished by listeners: Kawahara & Braver (2013) report that Japanese speakers, when presented with the appropriate orthography, can produce and distinguish at least 6 degrees of emphatic vowel length.[8]

Japanese short vowels also lengthen to satisfy a minimum word length requirement, namely, when they are the only vowel of a monomoraic word. While this might seem at first like a purely phonological process, in which the phonological length feature of the vowel is changed, Braver (2013) demonstrates that in the case the lengthening process is only near-neutralizing, such that lengthened short vowels having a shorter duration than long vowels in the same prosodic context.

In tone languages, lexical tones can have systematic durational differences, and these are used as perceptual cues (e.g. Liu & Samuel, 2004). Zhang (2000) argues that the complexity of contour tones is related to the duration of their realization in Mandarin. Flemming and Cho (2017) posit that contour tones have targets for the steepness of the rise in addition to initial and final pitch targets, which, taken together, are effectively tone-specific targets for duration, and that these can account for the degree of misalignment between tones and their segmental anchors.

[8] It is unclear to me how linguistically meaningful this ability is.

2.4. NLP models of duration

While much research has been done on individual factors contributing to duration, holistic, mathematical models of duration as a function of all of these factors in unison are harder to find in the academic literature (although some will be discussed on Chapter 3). However, in Natural Language Processing, and in particular in Speech Synthesis, such mathematical models are a practical necessity if any degree of prosodic naturalness is to be achieved in the synthetic voice. These, then, are examples of the first explicit proposals for holistic models of duration.

Klatt (1973, 1976) lays out the essentials of an algorithm for computing target durations for speech sounds that was ultimately used in the speech system DECtalk (previously KlattTalk or MITalk). It works as follows: each phoneme in the language has an inherent duration, as well as a minimum duration. The inherent duration is altered by a series of rules corresponding to individual lengthening or shortening effects. Each of these rules takes the amount by which the duration exceeds the minimum duration, and multiplies it by a constant K associated with that rule. The result after all of this is the duration target.

TABLE II. Rules for predicting vowel durations in strings of nonsense syllables spoken as a word in a carrier phrase.

Inherent phonological durations D_{inh} are derived from phrase-final monosyllables ending in a voiced stop, e.g., "bag" or "big."

Phone	D_{inh}	D_{min}	Ratio
/æ/	240	105	0.42
/ɪ/	160	65	0.42

Rule 1. If the postvocalic stop is voiceless, reduce the vowel duration by 45 msec.

$$D = D - 45.$$

Rule 2. Shorten a non-phrase-final vowel by about 35%, that is, set $K = 0.6$ in the equation below because D_{min} is about half of the inherent vowel duration

$$D = K * (D - D_{min}) + D_{min}.$$

Rule 3. Shorten an unstressed vowel by $K = 0.4$, except that a word-initial unstressed vowel of a polysyllabic word is only shortened by $K = 0.55$,

$$D = K * (D - D_{min}) + D_{min}.$$

Rule 4. Shorten all syllables in a polysyllabic word by 15%, that is, set $K = 0.78$.

$$D = K * (D - D_{min}) + D_{min}.$$

Figure 1: Table III from Klatt, 1976, p. 1217.

This algorithm amounts to a "multiplicative" or log-linear model of segment duration, except that what is calculated by the log-linear model is duration above a minimum duration threshold for that phoneme. If the minimum duration for each phoneme is known ahead of time, the appropriate constants K for each rule could be learned from duration data taken from a large speech corpus by fitting a basic multinomial linear regression with no interactions, where the dependent variable is the logarithm of the amount that observed segment durations exceed their minimum durations. Klatt, for his part, opts to adjust these parameters by hand until the results are subjectively satisfactory, probably due to the lack of available speech corpora with suitable coverage or rich enough prosodic annotation.

Van Santen et al. (1997) describe a language-independent procedure for developing generative models of duration, which was used by the Bell Labs Text-to-Speech system. After the phonological factors and classes of sounds relevant to the language in question have been selected, log duration is computed with a "sums of products" model: a linear model with many-way interactions. The overall shape of the model (essentially, what interactions should be included) needs to be determined by the researcher on a language-by-language basis, after which the parameters of the model are learned with a regression.

The main differences between the two models, aside from their approach to parameter estimation, are Klatt's use of a "minimum duration", and Van Santen et al.'s allowing for the possibility of interaction effects. However, both are essentially "template based" (Van Santen et al., 1997, pp. 228-30) log-linear models of duration, and therefore predict that the durational change of a segment undergoing two lengthening or shortening processes is in some sense proportional to the product of the changes that would occur if it underwent each of the processes individually. The truth of this prediction is an empirical question.

2.5. Interaction effects and the failure of linear models

While much is known about how various phonological factors influence duration individually, less is known about how they interact, and the workings of the grammar that determines durational targets from these factors. The applied models in the previous section assume that these durations are the result of multiplying coefficients together, each representing some such phonological factor.

Klatt (1973b) is one of the first (and perhaps only) attempts to explicitly investigate this very assumption. He devises an experiment to test the interaction between two factors already known

to affect the length of stressed vowels in English: the voicing of the following consonant (vowels are shorter before [t] than before [d]), and whether the vowel is part of a mono- or a disyllabic word (vowels are shorter in disyllabic words). He selects 40 pairs of monosyllabic and disyllabic words, such that the disyllabic word contains the phonological content of the monosyllabic word at its left edge (*need* / *needle*; *guess* / *guessing*; *room* / *rumor*; etc.).[9] He then computes duration averages for vowels in four contexts: in monosyllables with post-vocalic voicing (+V1), in disyllables with post-vocalic voicing (+V2), in monosyllables without post-vocalic voicing (-V1), and in disyllables without post-vocalic voicing (-V2). He assumes that the longest category, +V2, is the least marked case, and treats both effects as shortening effects. Since this experiment was conducted prior to the introduction of minimum duration into the model, his prediction was that the magnitude of two shortening effects individually, as determined by observing the durations of -V2 and +V1 as percentages of +V2, could be multiplied to predict the duration of -V1, the category to which both shortening effects had applied. In other words, by looking at the results for three of the four categories, Klatt's equation should predict the fourth.

[9] In the selection of these words there was no attempt to counterbalance the stimuli for vowel quality, or for the features (other than [voice]) of the consonants following the vowel. The relevant consonant was also sometimes [t] or [d], which was probably tapped in some of the disyllabic items, interfering with the supposed voicing distinction on the following consonant.

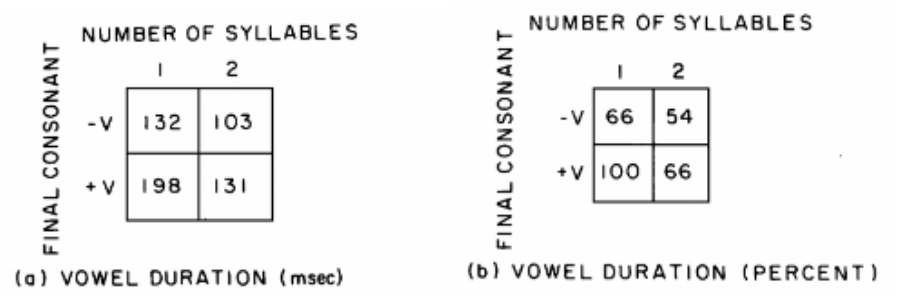

Figure 2: Interaction Results (Klatt 1973b, p. 1103).

This turned out not to be the case: the shortest tokens, vowels in disyllables before voiceless consonants, exhibited much less shortening than expected (in other words, there was a positive interaction effect). Klatt explains this result by positing that vowels are simply not infinitely compressible and therefore have a minimum duration below which they cannot easily be shortened further. The shortest category *would* have turned out as predicted, but ran against this maximum compressibility threshold. As discussed above, minimum durations were incorporated into his model and ultimately into the Klattalk text-to-speech system (Klatt, 1982), itself the basis for many of the systems that followed. While Klatt's stipulation about maximum compressibility is plausible as a reason for his model's over-predicting of the length of the shortest case, it is ad hoc in the context of this experiment: using inherent vowel duration, two effect sizes, and a minimum duration—a model with four features—any experimental results consisting of four data points can be fit perfectly. However, Klatt's fundamental question about how segment's durations are affected when they are affected by multiple shortening or lengthening processes, each of which is understood individually, is a crucial one.

A scattering of results related to interactions between intonational, prosodic, and segmental effects on duration can be found in the empirical literature, mostly from studies of moderately

sized corpora of reading-style speech. Some representative results are described here,[10] categorized by which factors were found to interact. For the most part, only results related to English are discussed.

2.5.1. Prosodic position x accent

Li & Post (2014), in a paper on L2 acquisition of rhythm, establish L1 baselines by collecting vowel duration data for native speakers, reporting average vowel durations in accented, unaccented, phrase-final, and phrase-medial positions.

	unaccented	accented
non-final	100% (baseline)	155.5%
final	162%	233.7%

Table 1: Mean English vowel duration by accentuation and phrase-finality reported by Li & Post (2014).

While in medial position, accented vowels are 55.5% longer than their unaccented counterparts, in final position, accented vowels are only 44.2% longer.

2.5.2. Prosodic position x lexical stress

In phrase-final two syllables words, the final syllable lengthens more when it is stressed, i.e. when the word is an iamb, than when it is stressless, i.e. when the word is a trochee (Turk & Shattuck-Hufnagel, 2007).

[10] See Fletcher (2010), section 2.2.2, for another overview.

2.5.3. Lexical stress x accent x vowel height / coda voicing

Word accentuation affects the length of both stressed and unstressed vowels, but its effects on stressed vowels are proportionally stronger—stressed vowels lengthen quite a bit due to accentuation while stressless vowels lengthen relatively less (Van Santen, 1992; Turk and White, 1999).

The difference in vowel length between vowels in primary stressed syllables and those in secondary stress syllables is greater for /æ/ than for /ɛ/ (De Jong, 2004).

The effect coda voicing on vowel duration is strongest in syllables with primary stress, less strong in syllables with secondary stress, and weakest (perhaps non-existent) in unstressed syllables (De Jong, 2004).

De Jong (2004) also finds three-way interactions between lexical stress, pitch accent, and vowel height, as well as between lexical stress, pitch accent, and coda voicing—all in the positive direction. In other words, stressed accented syllables are longer than predicted by model with only main effects, and this asymmetry is exaggerated for /æ/ as compared to /ɛ/, and for vowels preceding voiced obstruents as compared to voiceless.

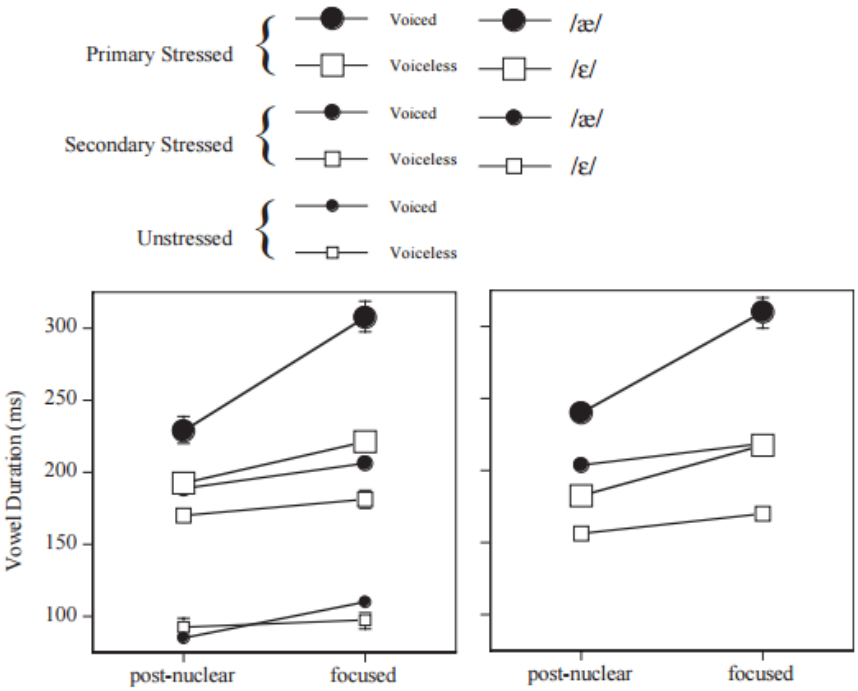

Figure 3: Figure 3 from de Jong, 2004 (p. 503). Average vowel durations for vowels before voiced and voiceless stops (left panel) and for /æ/ and /e/ (right panel). Plotted here are interactions between focus condition (x-axes) and stress (symbol size). Error bars indicate standard errors.

2.5.4. Accent x vowel height x coda voicing

Lengthening due to accentuation (nuclear pitch accent, as compared to pre-nuclear unaccented) is greater for /æ/ than for /ɛ/ (De Jong, 2004).

Lengthening due to accentuation is also greater for pre-voiced vowels than pre-voiceless ones (De Jong, 2004; Choi et al., 2016). Alternately stated, coda-voicing has a greater effect on vowel length in accented words than in unaccented ones—Choi et al. (2016) find that the effect of voicing is almost completely absent in deaccented words that occur before a focused word later in the phrase.

Choi et al. (2016) also find a significant *three-way* interaction between vowel height (/æ/ vs. /ɛ/), coda voicing, and accentuation, as shown in the leftmost (NAE) boxes in Figure 4.

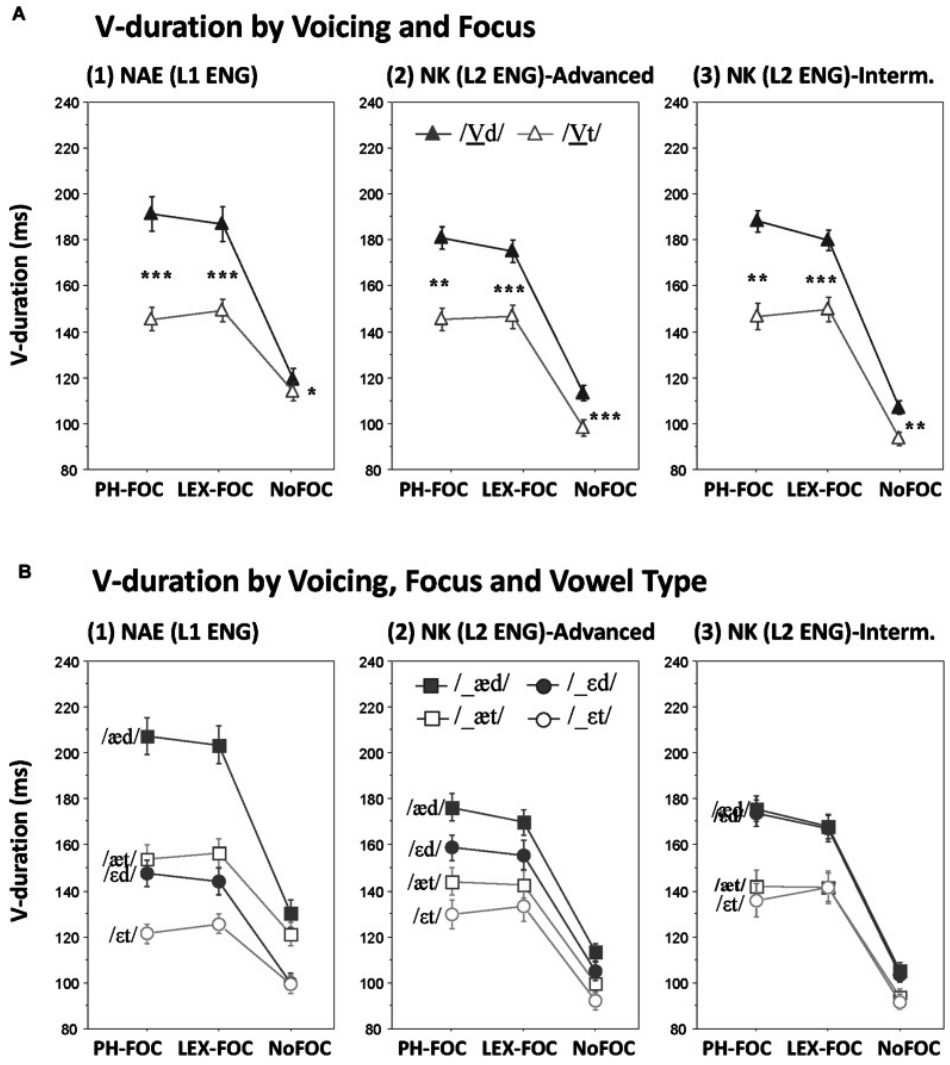

Figure 4: Figure 1 from Choi et al. (2016), p. 624. Effects of coda voicing on vowel duration. (A) Voicing × Focus interactions; (B) Voicing × Focus × Vowel type interactions, as produced by (1) native speakers of English, (2) Korean advanced learners of English, and (3) Korean intermediate learners of English (∗∗∗$p < 0.001$, ∗∗$p < 0.01$, ∗$p < 0.05$).

2.5.5. Prosodic position x coda voicing

Vowel lengthening in response to the segmental features of a following consonant (such as voicing or manner) is more prevalent phrase-finally than it is phrase-medially (Umeda, 1975; Cooper and Danley, 1981; Crystal & House, 1988), and, among phrase-medial vowels, more prevalent in word-final syllables (monosyllables and the stressed syllables of iambs) than in initial ones (Umeda, 1975), as well as in accented syllables compared to unaccented ones (Van Summers,

1987). Umeda (1975) finds that this interaction between segmental environment and phrase-finality holds even for reduced stressless vowels, with the effect of a following consonant on the length of schwa being much more evident in phrase-final position.

2.5.6. Vowel tenseness x coda voicing

Tense vowels lengthen more in response to a following voiced consonant than lax vowels do (Crystal & House, 1988).

2.5.7. Coda voicing x coda manner

Not only do both voicing and manner of coda consonants affect the length of the preceding vowel, but the voice value of stops matters more for vowel length than the voice value of fricatives (Crystal & House, 1988).

2.5.8. Phonemic length x syllable structure

English does not have contrastive length, but languages that do may show interactions between phonemic vowel length and other features influencing duration. For example, Broselow et al. (1997) examining cross-linguistic durational evidence for moraic structure, compare the effects of closed syllable shortening in a number of languages. They report that Arabic shows relatively little closed syllable shortening compared to other languages they examine, but of interest here is the fact that the closed syllable shortening effect was apparent for long /aː/, but not for short /a/, an interaction between the effects of syllable structure and phonological length.

2.6. An empirical generalization: hyperadditive lengthening

There is a surprising and hitherto unreported generalization hidden in these findings: nearly all of the reported interactions between duration-affecting phonological variables are positive, in

the sense that the segments undergoing multiple lengthening processes are *longer than expected*: already long segments are disproportionately susceptible to further lengthening. Put another way, if any pair of processes are taken to both have shortening effects on a segment (as in Klatt, 1973b), the interaction between them when both apply is such that the segment is *not as short as expected.*

The only clear example of a negative interaction, where a segment undergoing two lengthening effects turns out to be not as long as predicted (or where a segment undergoing two shortening effects is shorter than predicted), is the interaction between accentedness and phrase-finality seen in the data from Li & Post (2014). Interestingly, both of these factors relate in some way to intonation, or at least to phrase-level prosody.[11]

[11] An additional potential exception is the interaction between coda voicing and coda manner (stop vs. fricative) on vowel duration; however, the existence and direction of the main effect of obstruent manner on vowel duration is debatable (see discussion in Chrystal & House, 1988). Because one of the main effects is putative, it is unclear whether the interaction reported by those authors is consistent with the present generalization.

	vowel features	coda features	coda complexity	lexical stress	accent	phrasal position
two-way interactions						
word-length (syllables)		Klatt (1975)				
vowel features		Crystal & House, 1988; Choi et al., 2016		De Jong, 2004	De Jong, 2004; Choi et al., 2016;	
coda features		Crystal & House, 1988	Katz (2010)	De Jong, 2004	De Jong, 2004; Choi et al., 2016	Umeda, 1975; Cooper & Danley, 1981; Crystal & House
coda complexity						
lexical stress					De Jong, 2004; Van Santen, 1992; Turk and White, 1999	Turk & Shattuck-Hufnagel, 2007
accent						Li & Post, 2014
three-way interactions						
lexical stress × accent	De Jong, 2004	De Jong, 2004				
vowel features × accent		Choi et al., 2016;				

Table 2: Reported "interactions" between effects between factors affecting vowel duration. All interactions were in the positive direction (if the effects are treated as both being lengthening or both being shortening effects), except for the interaction between pitch accent and phrasal position, and potentially the interaction between coda manner and coda voicing.

This apparent asymmetry is perhaps unexpected, since negative interacts seems just as plausible as positive ones *a priori*, and could just as easily be fit by any linear or log-linear model that includes interaction effects, such as those proposed by Van Santen et al. (1997).

I will henceforth refer to this empirical pattern as the "Hyperadditive Lengthening Generalization," though it could just as well be named the "Hypoadditive Shortening

Generalization," depending on whether the effects interacting are taken to be lengthening or shortening effects. It is stated formally as follows:

The Hyperadditive Lengthening Generalization

The duration of a segment taken to be undergoing both of two lengthening processes will be longer than the duration predicted by a multiplicative model, given the magnitude of the processes when they apply individually, and using the case where neither applies as the baseline. Equivalently, the duration of a segment taken to be undergoing both of two shortening processes will not be as short as predicted.

In Chapter 5, further experimentation will corroborate a number of these interactions. In the discussion section of that chapter, potential mechanisms for deriving the Hyperadditive Lengthening / Hypoadditive Shortening pattern will be explored, including an explanation couched in the framework of phonetic harmonic grammars with asymmetric constraints developed in Chapter 4.

2.7. The task at hand

Any holistic model of the phonetic component of the language apparatus needs to include a way for this very diverse set of factors, ranging from phonetics, to segmental phonology, to prosodic, metrical, and intonational phonology, to speech rate, to be a part of the computation of duration and timing in speech. More importantly, it must make a claim about the mechanism by which these factors can and cannot interact, in a way that is consistent with our (thus far limited) empirical knowledge about such interactions.

The use of log-linear models for computing duration, in which each factor is essentially a multiplier on the duration[12] of the segment(s) it affects, is ubiquitous in the speech technology literature, possibly due to their simplicity and implementability. This choice, however, is not theoretically motivated, and may not even be observationally adequate (Klatt, 1973b).

The ensuing chapters take an approach in line with an emerging framework that one might call "Harmonic Phonetics", following recent authors (Flemming, 2001; Braver, 2013; Windmann et al., 2015; Flemming & Cho, in print) in employing harmonic constraint grammars, already widely in use by phonologists, to map from phonological representations to phonetic ones. These models naturally accommodate multiple, potentially competing constraints on the durations of segments and of larger constituents, predict interesting nonlinearities as a necessary outcome of the constraint-based nature of the model itself, and come with a well-understood algorithm for learning the model parameters (constraint weights) from example data.

[12] Or the duration above a baseline, in the case of Klatt (1973, 1976)

3. Phonetic constraint grammars

There are several existing proposals for adapting or expanding varieties of constraint grammars in common use by phonologists (in particular, Classical Optimality Theory and Harmonic Grammar), to the domain of phonetics. In addition to adapting different flavors of optimality theory, these proposals also differ significantly in their formulation of phonetic constraints, particularly in how they assign violations to candidates, and in how GEN and EVAL are treated.

Phonetic constraints differ crucially from the kinds of constraints employed in phonology in that the linguistic objects they penalize are articulatory or acoustic representations which contain real-number values, rather than symbols which fall into discrete categories. Relatedly, while phonological constraints tend to be formulated so as to assign to any particular candidate an integer number of violations corresponding to how many pieces of the candidate's structural description violates the constraint, it is less obvious how a constraint on, say, tongue height, or F1, should assign violations to a candidate pronunciation with a phonetic value that differs from what the constraint would consider optimal.

There is also the question of where phonetic constraints fit into the overall grammar, and how they interface with the phonological component, if these components are in fact separable. A theoretical choice must be made about what linguistic representations phonetic constraints should map from, and what they map to.

Prior authors have taken a variety of approaches, differing on all of these issues. This chapter outlines several existing proposals, and concludes with a summary or "taxonomy" of the various, pseudo-independent theoretical and implementational choices that need to be made in formulating

phonetic constraint grammars, where the existing proposals fall in this taxonomy, and what other combinations might be possible.

3.1. Proposing phonetic constraints

One of the first authors to propose that constraint grammars be used in phonetics was Zsiga (2000), who did so in the context of investigating empirical patterns related to the alignment and degree of overlap in adjacent consonants at word-boundaries in Russian and English. Zsiga finds that the two languages behave phonetically differently given the same phonological conditions, both in the way consonants align and in the kinds of phonetic variation that can occur with respect to these alignments. Zsiga proposes that there must indeed be a phonetic grammar, separate from the phonological grammar, and that the best way to model her data in particular is with language-specific phonetic alignment constraints. These constraints are analogous to alignment constraints in phonology, but which govern phonetic-level phenomena like consonant overlap and coarticulation, and contain real-number phonetic values. The particular alignment constraints proposed are consistent with the "articulatory window" view of phonetics: an anchor point in one consonant is required by a constraint to lie within some temporal range defined in relation to an adjacent consonant, but there can be free variation within this range, or variation in response to other phonological factors.

Because phonetic values are subject to multiple pressures, and the resulting pattern is often one of compromise, Zsiga argues that the strict dominance found in many varieties of OT is not as well-suited to capturing patterns of phonetic variation as constraint weighting, which allows for compromise candidates to win out over more extreme candidates.

3.2. Classical OT with categorical constraints

Boersma (2009, 2011) employs "cue constraints" (2009) as part of a larger framework, Bidirectional OT (2008, 2011), a grammatical architecture in which various levels of phonetic and phonological representations are mapped to and from each other using a series of ranked constraint grammars. These same grammars are used for both perception and production (thus the "bidirectionality"). In the phonological levels of this framework, underlying forms are mapped to and from surface forms by way of ranked faithfulness and markedness constraints, as in classical OT. However, surface forms are also mapped to auditory forms by another level of the grammar, consisting only of ranked cue constraints. Yet another level, which is relevant only for production and not for perception, maps from auditory forms to articulatory forms, which are then interpreted by the motor system.

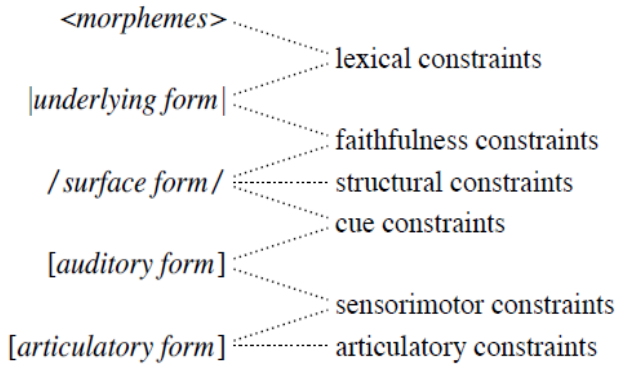

Figure 5: Figure 2 from Boersma 2009, p. 60.

Auditory forms are here essentially sequences of distinct auditory cues, such as periodic vibration, silence, or aperiodic noise. Describing the process by which a Russian word, [.tak.], would be perceived and eventually loan-adapted as /.ta.ku./ by Japanese speakers, Boersma (2009) gives the following example:

"In a narrower transcription, the auditory form is [_^{t}a_^{k}]: as you can see in a spectrogram, this sound consists of (at least) a silence ([_]), followed by a high frequency brief noise ("burst") ([^{t}]), followed by a loud periodic ("sonorant") sound with formants around 1000 Hz ([a]), followed by another silence, followed by a burst with a peak around 2500 Hz ([^{k}])." (p. 10)

Cue constraints, therefore, govern the relationship between a surface form which is a sequence of symbols (each standing for a bundle of binary phonological features), and an auditory form which is also for the most part a sequence of symbols standing for acoustic features such as silence, noise, periodic sound, but which crucially also includes parameters, such as formant frequency (and presumably other acoustic variables such as duration and intensity), that have associated real-number values in units like Hz.

[380 Hz]	*/a/ 320 Hz	*/a/ 380 Hz	*/i/ 460 Hz	*/e/ 320 Hz	*/a/ 460 Hz	*/i/ 380 Hz	*/e/ 380 Hz	*/i/ 320 Hz	*/e/ 460 Hz
/a/		*!							
☞ /e/							*		
/i/						*!			

Figure 6: Boersma 2009, p. 30

Figure 6 demonstrates how perception of a vowel would work using cue constraints. Because the tableau is demonstrating perception, the input is an acoustic representation, and the candidates are potential surface representations. Since Boersma's grammars are bidirectional, the same cue constraints with the same ranking are used for production, so we can easily imagine a similar tableau for vowel production, with a surface representation input and very many acoustic representation outputs, perhaps one for each minimally distinguishable F1 value in Hz.

There are several things worth noting here. The first is that violations in this theory are categorical, just as in the other levels of the grammar. The second is that, wherever real-number values such as formant frequency are concerned, very large families of constraints will need to be present, such as *\/i/ 460 Hz, *\/i/ 380 Hz, and *\/i/ 320 Hz, identical to each other except for the numerical targets they penalize. Thus, in even a production grammar for determining a single formant of a vowel in isolation has very many constraints,[13] and the formant values associated with a particular vowel category result from their ranking.

While the evaluation function EVAL works much the same way as in Classical OT, the nature of GEN depends on the directionality of the grammar. For perception, since the cue constraints are mapping acoustic to surface representations, it enumerates possible SRs much as in the phonological component, but for production it must generate all possible acoustic representations, including ones which differ only in numerical values such as formant frequency.

3.3. Harmonic grammar with gradiently violable constraints

Flemming (2001), followed by others (Katz, 2010; Braver, 2013; Flemming & Cho, 2017), propose phonetic constraint grammars of a very different sort, wherein a single constraint associated with a phonetic target can penalize a candidate gradiently as a function of how far off that candidate is from the target. These authors employ this framework in modeling a number of phonetic and/or phonological phenomena, including duration and timing (Katz, 2010).

[13] The exact number of constraints in each family is presumably constrained by limitations on the precision of our perceptual and articulatory system. While constraints may not need to be spaced as closely as 1Hz apart, the spacing cannot be much coarser if we are to explain any empirical data where relatively small changes to acoustic cues can cause category differences, or where phonological factors can have rather small effects on articulatory targets.

Windmann et al. (2015) similarly employ gradiently violable constraints, albeit with very different violation functions and without appealing to phonetic targets (or, equivalently for some of the functions, fixing this target at zero), in their own model of speech timing.

3.3.1. Flemming's parabolic constraints

Flemming's larger program is a push towards a unified model of phonetics and phonology, in which underlying forms are mapped directly to acoustic ones, using the same constraints to explain categorical "phonological" phenomena (such as assimilation) and gradient "phonetic" ones (such as coarticulation). However, the kinds of constraints he proposes work equally well as a model of just the phonetic component of the grammar, on the more conservative view that this is distinct from the phonological component and takes phonologically derived surface representations as its input, as is assumed by Zsiga (2000) and Boersma (2009).

Two key innovations define this class of phonetic grammar. The first is that it is an implementation of Harmonic Grammar (Legendre et al., 1990), wherein constraints are weighted rather than ranked, and the winning candidate is the most harmonic. The second is that phonetic constraints can have in their formulation numerical phonetic targets (such as a specific target value for f2 for some vowel), and are gradiently violable. In particular, candidates subject to a constraint incur violations proportional to the square of the distance between the phonetic value of the candidate and the target phonetic value specified by the constraint.

To give the simplest possible example, we could imagine a grammar with two constraints on some phonetic value (say, vowel duration). Constraint 1 has a target T_1 and weight w_1, and the cost of violating this constraint $C_1(x) = w_1(x - T_1)^2$ where x is the candidate phonetic value. Note that

when $x = T_1$, no violations are incurred. Constraint 2 is identical, but has its own target T_2 and weight w_2.

The total cost C incurred by a candidate phonetic value x in this grammar is simply the sum of the costs of the two constraints, given in (1), and visualized in Figure 7.

(1) $$C(x) = w_1(x - T_1)^2 + w_2(x - T_2)^2$$

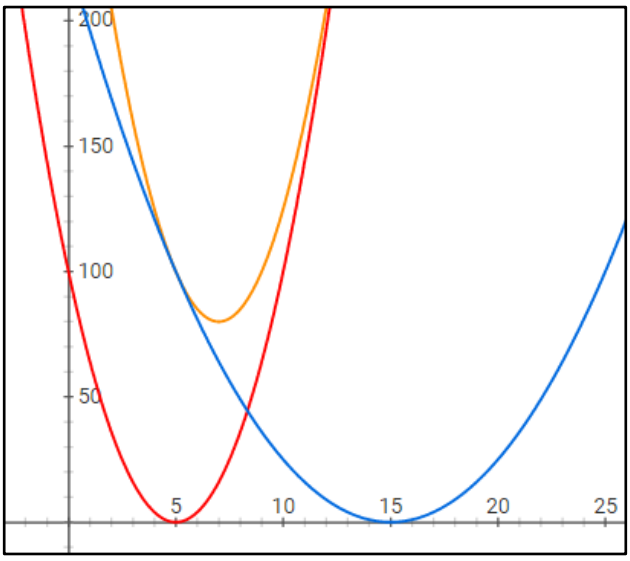

Figure 7: The sum of two constraints' violation functions. Red: cost of violating C1(x) where T1 = 5, w1 = 4. Blue: cost of violating C2(x) where (T2 = 15, w2 = 1). Orange: the total cost C(x). Note that the minimum of the total cost function does not lie at the target of either constraint, but in between the two targets.

The most harmonic phonetic value for x can then be found by taking the derivative of C with respect to x, setting this derivative to zero (since the first derivative is zero when the cost function is minimized), and solving for x.

(2) $$C'(x) = 2w_1(x - T_1) + 2w_2(x - T_2) = 0$$

(3) $$x = \frac{w_1T_1 + w_2T_2}{w_1 + w_2}$$

Note that the most harmonic value will always be a compromise between the two constraint targets (unless one of the constraint weights is zero), and the distance of this value from a constraint's target will be smaller if that constraint has a larger weight, as we might expect. In fact, the ratio of distances between the optimal value and the two constraint targets is exactly the inverse of the ratio of the weights of the constraints, as can be seen in (4).

(4) $\frac{x-T_1}{T_2-x} = \frac{w_2}{w_1}$

The situation becomes slightly more complicated when multiple phonetic values are selected by the grammar in parallel. In an example given by Flemming (2001), coarticulation is modeled as adjustment of the F2 of a vowel and the F2 locus of an adjacent consonant in order to better satisfy a constraint MINIMISEEFFORT, which assigns violations proportional to the square of the F2 distance between adjacent sounds in CV sequences, penalizing large formant transitions. Competing with this constraint are two IDENT constraints, which require vowel formants and consonant loci to match predetermined targets (T and L, respectively) which correspond to the most "faithful" renditions of these sounds.[14] Since Harmonic Grammar is being used, constraint violations are multiplied by the weight of the constraint. The costs of violating each of these three constraints are summarized in Table 3.

[14] There is some question here as to where exactly these targets / loci should reside. In order for Flemming's IDENT constraints to literally be faithfulness constraints, the targets would need to be part of the lexical representations that are inputs to the grammar, which would introduce numerical values into the representations of individual lexical items. Flemming instead posits something like a phoneme inventory (or inventories, since contrast is contextually limited) to be a part of the grammar, and for these inventories to contain the phonetic targets, which are themselves subject to meta-constraints on entire systems of contrast which enforce dispersion (Flemming, 2004). In the chapters that follow, I will take an alternate approach, taking these targets to simply be parameters of the "faithfulness" constraints themselves, much like their weights.

	Constraint	*Cost of violation*
IDENT(C)	$F2(C) = L$	$w_c(F2(C) - L)^2$
IDENT(V)	$F2(V) = T$	$w_v(F2(V) - T)^2$
MINIMISEEFFORT	$F2(C) = F2(V)$	$w_e(F2(C) - F2(V))^2$

Table 3: (11) from Flemming, 2001, p. 19 : Cost functions for the three constraints.

Once again, as demonstrated in the following tableau, candidate values for the beginning and end of an F2 transition in a CV sequence which compromise among the three constraints, violating each of them to some degree, will outperform values which fully obey any one constraint to the detriment of the others.

F2(C)	*F2(V)*	IDENT(C)	IDENT(V)	MINEFFORT	*total cost*
1700	1000	0	0	490,000	490,000
1500	1200	40,000	40,000	90,000	170,000
1350	1350	122,500	122,500	0	245,000

Table 4: Table I from Flemming, 2001, p. 21. Evaluation of example candidate values for F2(C) and F2(V), with L = 1700 Hz, T = 1000 Hz, and all weights set to 1.

The total cost is here a function of two phonetic values in the candidate, F2(V) and F2(C), given in (5). The cost function is also visualized in Figure 8, where it can be seen to have a bowl-like shape, with a global minimum.

(5) $C = w_c(F2(C) - L)^2 + w_v(F2(V) - T)^2 + w_e(F2(C) - F2(V))^2$ (p. 20)

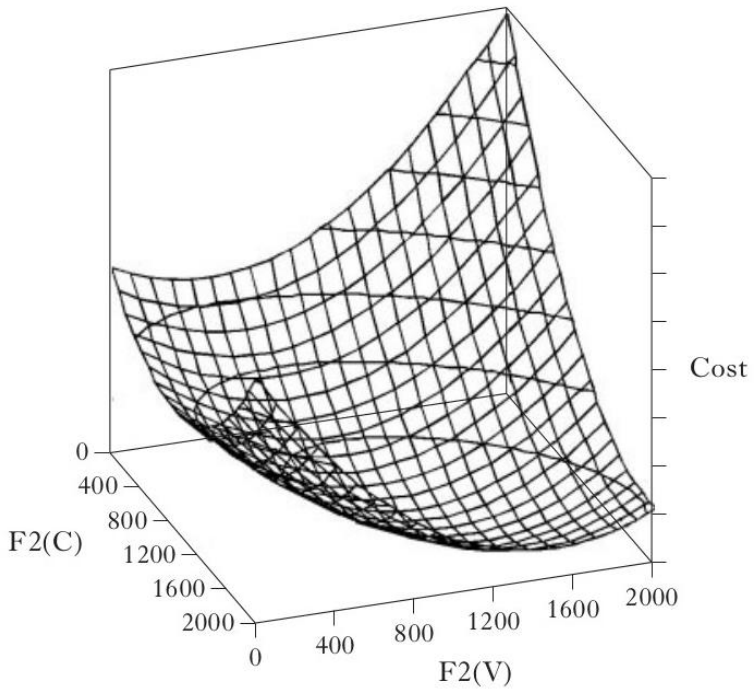

Figure 8: Figure 3 from Flemming, 2001, p. 21. Cost plotted against F2(C) and F2(V). The minimum is located at F2(V) = 1233 Hz, F2(C) = 1467 Hz.

The maximally harmonic values for F2(V) and F2(C), i.e. the global minimum of the bowl in Figure 8, can once again be found with calculus, but in this case this requires taking *partial* derivatives of the cost function with respect to F2(V) and F2(C), setting both partial derivatives to zero, and solving the resulting system of equations. The solutions to these equations, which are the maximally harmonic phonetic values given this grammar fragment, are given in (6-7).

(6) $F2(C) = u_c(L - T) + L$ where $u_c = \frac{w_e w_v}{w_e w_c + w_v w_c + w_e w_v}$

(7) $F2(V) = u_v(L - T) + T$ where $u_v = \frac{w_e w_c}{w_e w_c + w_v w_c + w_e w_v}$ (p. 22)

3.3.2. A note on GEN and EVAL

It is here worth taking a moment to ponder what GEN and EVAL are like under this account. While the tableau in Table 4 includes only a few choice candidates selected for the purpose of

illustration, the actual candidate set is assumed to be continuous: all real-number values for all the acoustic features being determined are possible outputs of the grammar. If two acoustic features are being selected, the candidate set comprises at the very least[15] a 2-dimensional vector space, each candidate being represented by a vector composed of its values for these two features. This makes GEN in these grammars different from GEN in Classical OT, as the number of candidates is uncountably infinite, even when only considering candidates which have the same general phonetic shape (what would traditionally be called a Surface Representation).

EVAL, similarly, is generally construed in optimality theoretic accounts as enumerating the candidates or a subset thereof, and (in the case of HG) selecting the most harmonic. However, as pointed out, the points in a vector space are not enumerable. Flemming's method for implementing EVAL—computing the most harmonic phonetic values directly—involves differentiation and, when multiple phonetic values are involved, linear algebra (in that it involves solving the system of equations that results from taking multiple partial derivatives). While this method demonstrates how a winner can be found for these simple examples, for more sizable constraint grammars and for candidates in which multiple acoustic values are computed in parallel, this step can become computationally very complex as compared to the traditional mechanism for EVAL, especially since the algebraic form of the harmony equation is not the same in all cases, but rather depends on the shape of the input and of the constraints, the number of phonetic values being derived, their relationship to each other in the input, and so on. This concern about implementability, not only

[15] For Flemming, since these grammars must also govern alternations which are more traditionally considered phonological (deletion, epenthesis, changes in phonetic category, metathesis, stress alternations, and so on), the sequence of acoustic events also needs to be selected, in addition to these events' exact phonetic targets. The candidate set is presumably therefore all the points in the various vector spaces projected by all possible sequences of phonetic events, and not just one vector space.

for human speakers and learners of these grammars but even for linguists interested in algorithmically learning the parameters of such grammars from data, will be discussed again in the immediately following chapter.

3.3.3. Parabolic constraints on duration and timing

Katz (2010) employs similar constraint grammars to model the data from an empirical study of "compression" effects in English—primarily compensatory vowel shortening in response to combinations of adjacent onset and coda consonants—as a function of onset and coda complexity and of the properties of the specific consonants involved.

As might be expected by the name "compression", the idea is that segments have some optimal duration, but that larger constituents such as syllables do as well, and when the segments in a syllable would at their optimal durations together exceed the preferred length for a syllable, one or more of them has to shorten, or the syllable must have a duration longer than optimal, or both. This intuition can be captured straightforwardly with parabolic phonetic constraints on the durations of syllables and of the segments therein, such that the maximally harmonic candidate involves a compromise—the result of "trying to fit partially-malleable objects into a partially-malleable container" (Katz, 2010, p. 91).

The prediction of such an account is that as more and more material is added to a syllable, the segments in it should all continue to shorten, though some might do so less than others due to the weights on the segmental constraints. Empirically, looking at compression of the vowel as consonants are added to the onset and coda, Katz finds that the addition of a simplex onset or coda does result in compensatory shortening (of a "simplex" sort, i.e. compared to a similar syllable in which the onset or coda is empty), but that the degree depends on the consonant involved. The

concatenation of additional consonants to the periphery of the syllable, creating complex onsets or codas, sometimes but not always induces further "incremental" compensatory shortening of the vowel, depending on the sonority of the consonant closest to the vowel, and whether the cluster is in onset or coda position. Katz finds that "liquids condition incremental CS in both onset and coda position, nasals do so only in onset position, and obstruents don't clearly induce incremental CS in onset or in coda position." He also finds that "the amount of incremental CS for items with liquids as the inner consonant appears to be greater in coda than in onset position, especially for /l/" (p. 90).

A second issue arises with respect to segmentation of the data (a necessary step if durations are to be measured). As anyone who has annotated phonetic production data can attest, there are rarely clear boundaries between segments, especially between vowels and sonorant consonants, and substantial parts of the duration of the speech signal are best described as transitions rather than steady states which clearly belong to one segment only. Rather than taking the approach of placing segment boundaries in the middle of these transitions, Katz uses a finer grained segmentation in which both steady states and transitions are coded for, and in fact posits that these transitions and their perceptual properties might help to explain some of the asymmetries seen in the data.

Katz posits that the constraints on segment duration are fundamentally constraints on the recoverability of acoustic cues for the segment. These cues are taken to present in the segment itself, but due to coarticulation, can also be present in segmental transitions and even in adjacent segments. The presence and quality of such cues, however, depend on the segments. Therefore, Katz takes the recoverability to be equal to the duration of the steady state of the segment times some coefficient *i*, plus the duration of an adjacent segment transition multiplied by some

coefficient *j*, plus the duration of the adjacent segment times some constant *k*, where *i*, *j* and *k* depend on the acoustic and coarticulatory properties of the segments, and represent the extent to which the steady state, transition, and adjacent segment contain cues for the segment in question.

Katz's model, in summary, contains two constraint types, one on syllable duration (C_1), and one on segment recoverability (C_2), defined in (8) and (9).

(8) $C_1 = w_1 \bullet (t_\sigma - d_\sigma)^2$ (p. 98)

(9) $C_2 = w_2 \bullet (t_s – (i\ d_s + j\ d_t + k\ d_a))^2$ (p. 100)

…where w_1 and w_2 are the constraint weights, t_σ and t_s, are the target durations for the syllable and the segment, d_σ, d_s, and d_a are the candidate durations of the syllable, segment, and segment transition.

The combined cost of a candidate with just a consonant *x* and an adjacent vowel *y* is therefore the sum of the costs incurred by the syllable duration constraint, which involves the total duration (the duration of the consonant, vowel and transition), and two copies of the segment recoverability constraint: one for the consonant and one for the vowel (following Flemming (2001), Katz considers these to be in some sense the same constraint, so they share a weight, w_2).

(10) total cost $= w_1 \bullet ((d_x + d_t + d_y) – t_\sigma)^2 + w_2 \bullet (n\ d_y + m\ d_t + l\ d_x) – t_x)^2 + w_2 \bullet ((k\ d_x + j\ d_t + i\ d_y) – t_y)^2$ (p. 101)

Since it is primarily vowel duration that is being investigated, Katz makes several simplifications for the purposes of illustration. One is to assume that consonant cues are only present in the steady state of the consonant. Another simplification is that the baseline recoverability coefficient of the steady state of the vowel, *i*, is equal to 1—this does no harm, since what is important is the ratio between *i*, *j*, and *k*. These two simplifications reduce the equation to (11) below. Lastly, the duration of the transition is held constant, and assumed not to vary (though

it of course does systematically vary in Katz' data). These simplifications are made so the predictions of the model with regards to compensatory shortening effects on the duration of the vowel proper in response to syllable complexity can be feasibly investigated.

(11) $\text{total cost} = w_1 \bullet ((d_x + d_t + d_y) - t_\sigma)^2 + w_2 \bullet (d_x - t_x)^2 + w_2 \bullet ((k\, d_x + j\, d_t + d_y) - t_y)^2$ (p. 102)

Once weights for the constraints and recoverability coefficients have been selected by hand, along with the fixed transition duration, the result is a grammar that can assign a cost to any candidate consisting of just a pair of durations—one for the steady state of the consonant and one for that of the vowel. As in Flemming (2001), the winning candidate can then be found by calculating the durations for these segments which minimize the total cost function in (11), taking partial derivatives and solving the resulting system of equations. For reasons already discussed, as long as none of the weights are set to zero, the winning candidate will generally be one which obeys none of the constraints entirely, but one in which the segments have each adjusted in duration as needed to create a syllable duration preferred by the syllable level constraint, thus accounting for the general pattern of compensatory shortening.

The differences between adjacent consonants in their propensity to induce simplex (i.e. not incremental) compensatory shortening on the vowel can now be explained as well: if the coefficient on the recoverability of the vowel from the transition, *j*, is set to a higher value, the optimal duration of the steady state of the vowel predicted by the grammar will reduce. This is because, since the transition contains better vowel cues, the vowel duration constraint can be satisfied with less duration in the steady state portion of the vowel. The same is true of the coefficient on the recoverability of the vowel from the adjacent consonant, *k*. This prediction closely matches Katz' empirical findings: the consonants that are *a priori* expected to contain

better vowel cues induce less (simplex) compensatory shortening. In a later chapter, Katz confirms experimentally that the consonants which induce more shortening do in fact contain better vowel cues.

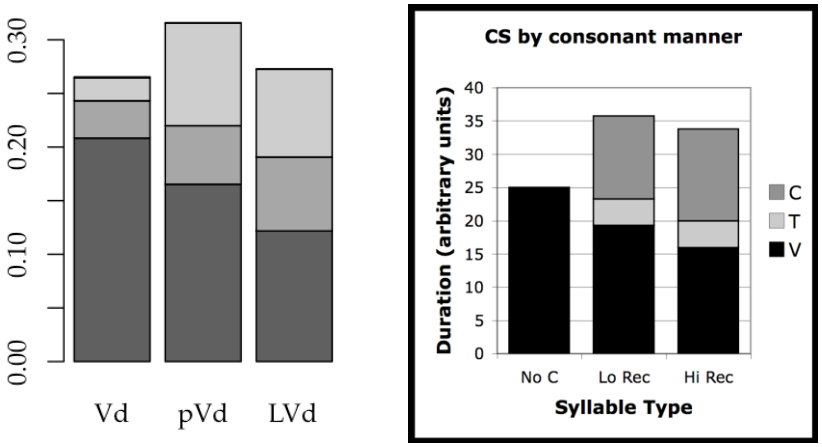

Figure 9: Figure 3.4. from Katz, 2010 (p. 112). Data from the production experiment (left) and model predictions (right) for consonant manners with high (rightmost bars) and low (center bars) vowel-recoverability coefficients. For production data, durations are in seconds. The upper bars for vowel-initial items represent closure and transition durations, in realizations where these categories are applicable.

The differences between consonants in their propensity to induce incremental compensatory vowel shortening is also framed as a result of their perceptual properties, but the failure of some types of consonants to induce any incremental shortening at all is unexpected given the constraints: because of pressure from the constraint syllable duration, additional crowding should result in additional compression. To remedy this situation, Katz appeals to floor effects on the duration of vowels, positing that since vowels are not arbitrarily compressible, incremental shortening has its limits. Interestingly, this is same solution proposed by Klatt (1973b) in response to the same problem, namely that Klatt's (completely different) duration model was also over-predicting the degree to which vowel shortening should occur in the cases where shortening effects should be the strongest. This pattern of hypo-additive shortening is in fact consistent with the Hyperadditive Lengthening Generalization described in Chapter 2. Explanations for such patterns which do not

rely on maximum compressibility effects will be hinted at in Chapter 4, and in the discussion section of Chapter 5.

Another author who uses parabolic constraints on duration is Braver (2013), who does so in the context of investigating a vowel lengthening process in Japanese wherein phonologically short vowels are lengthened in monomoraic words, presumably in response to a phonological requirement that words be at least bimoraic. As Braver demonstrates, short vowels lengthened in this way do not become as long as underlyingly bimoraic long vowels, resulting in near-neutralization of length in these environments.

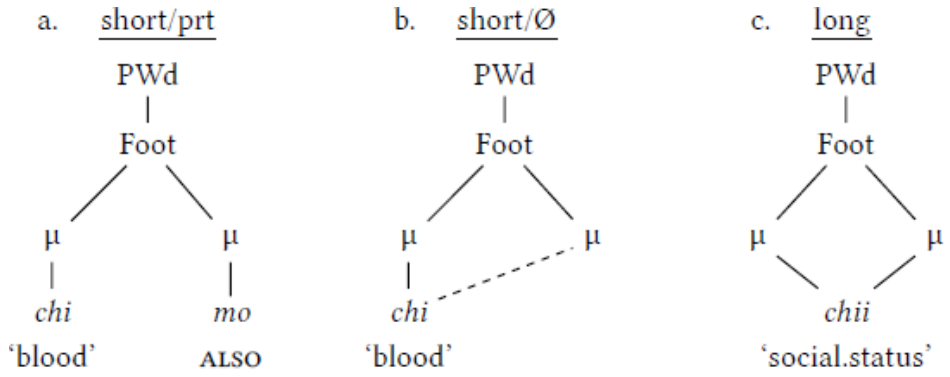

Figure 10: (8) from Braver, 2013 (p. 127). Three degrees of phonetic vowel length in Japanese.

In order to model his duration data, Braver also uses a weighted constraint grammar with the parabolic durational constraints from Flemming, 2001, with which the reader is by now familiar. However, his grammars make use of a phonetic version of Output-Output correspondence, or "transderivational identity." In particular, he employs a constraint (12) which governs the similarly of the phonetic durations of corresponding vowels in members of a morphological paradigm: the durations of vowels in less frequent paradigm members should, according to these constraints, not be too far off from the durations observed in the more frequent paradigm members.

(12) OO-ID-DUR: “The duration of a segment in the candidate should be faithful to the duration of the same segment in the base—the most frequent type in the candidate’s inflectional paradigm as applied to the candidate’s root.” (p. 134).

Because monomoraic noun stems are in Japanese most often followed by function words called “particles,” which form a prosodic word with the stem to which they attach, these stems generally find themselves in bimoraic prosodic words such that no lengthening is necessary, and that it is only in the comparatively rarer case that these words are not followed by a particle (such as when they are spoken in isolation) that an additional mora must be inserted, and lengthening occurs. In these cases, where underlyingly short vowels are associated with an additional mora due to the bimoraic word requirement, OO-ID-DUR would nevertheless prefer that they not lengthen, since this would result in phonetic paradigm non-uniformity compared to the more frequent tokens of these words where the vowel is associated only to one mora. In competition with OO-ID-DUR are more general constraints ((13)-(14) which govern the relationship between morae and duration, ensuring that long vowels are in fact longer in the language.

(13) DUR(μ) = TARGETDUR(μ): “The duration of a mora-bearing unit in the candidate, which bears a single mora in the output, should match the target (canonical) output duration of that mora-bearing unit (when it bears one mora) in the language at large.” (p. 129-30).

(14) DUR(μμ)=TARGETDUR(μμ): “The duration of a mora-bearing unit in the candidate, which bears two moras in the output, should match the target (canonical) output duration of that mora-bearing unit (when it bears two moras) in the language at large.” (p. 131).

While in the simpler cases (short vowels corresponding to a single mora and long vowels corresponding to two) vowel duration is straightforwardly governed by one of the DUR constraints, in the lengthening case there is competition between OO-ID-DUR, which prefers a shorter vowel, and D(μμ)=TARGETDUR(μμ), which prefers a longer one. Since these constraints use the (by now familiar) parabolic violation functions based on deviance from durational targets, the winner in these cases will show compromise between short and long. The DUR=TARGETDUR constraints are taken, however, to have a higher weight, such that derived lengths will be closer to those of the underlying long vowel category. In this way, near-neutralization is explained.

Flemming & Cho (2015) use the Flemming (2001) framework to provide an account of the phonetic realization of both the F0 and timing aspects of the rising tone in Mandarin across multiple speech rates. Prior authors had variously described the phonetic realization of this contour tone as being determined by the alignment of the endpoints of the rise to segmental anchor points (A_L and A_H), the slope of the rise (T_s), and the magnitude of the rise (T_M), with different authors employing different subsets thereof in their models of tone. The four specifications mentioned here cannot all be specified independently, because any three of them predict the fourth, as illustrated in Figure 11.

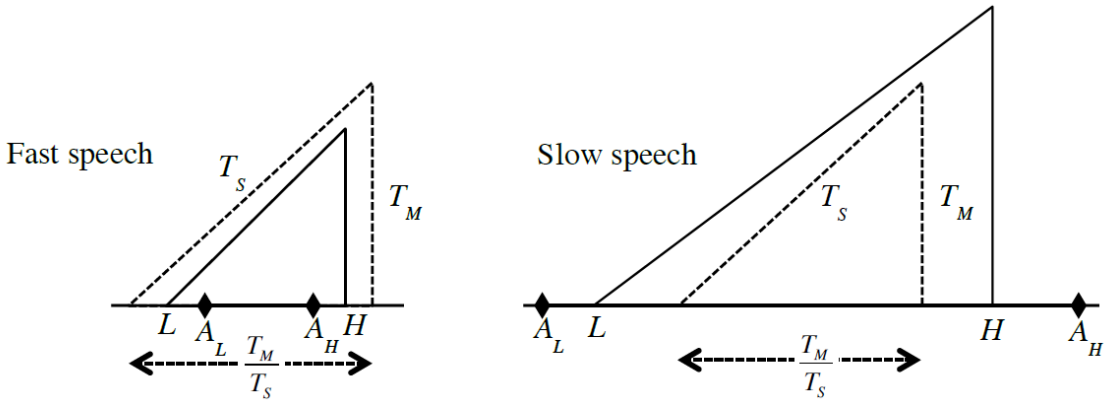

Figure 11: Figure 12 from Flemming & Cho 2017 (p. 20): "Schematic illustration of the conflict between realizing the magnitude, slope, and alignment targets for a rising tone. The dashed lines show the shape of the rise that satisfies the targets for rise magnitude and slope, while the solid lines schematize the actual slope and magnitude of the rise appropriate for the illustrated intervals between the alignment targets A_L and A_H."

Flemming and Cho therefore first empirically investigate which of these properties might be invariant across speech rate. Surprisingly, they find that in fact none of them are invariant, and that instead, as segmental material is made longer or shorter by changes in speech rate, all four properties vary. In particular, as speech rate increases, the beginning and endpoints of the rise come further before and after their segmental anchors, respectively, the slope increases, and the overall magnitude decreases, all compensating in tandem for the shortened duration of the segmental material.

The authors suggest that this is because all four of these specifications (the two alignment anchors, slope, and magnitude) are constrained by the grammar, in that speakers are aware of optimal values for each, such that realization of the tones is phonetically over-specified. Since the four tonal specifications cannot all be satisfied absolutely, actual realizations will involve compromise. To model their data, they propose four phonetic constraints, with targets and parabolic violation functions, which again turn out to be best satisfied by compromise candidates in cases where the constraints cannot all be satisfied. The authors take the empirical facts in this

case, namely the lack of phonetic invariants in contour tone specification, to support their phonetic HG framework.

3.3.4. Targetless constraints with heterogeneous violation functions

Windmann et al. (2015) propose a model for duration which also uses the Harmonic Grammar formalism, and phonetic constraints on duration with continuous violation functions. However, the violation functions used are very different from those already discussed, and the shapes of these functions are constraint-specific. In particular, the authors treat realization of segment duration as arising from the compromise between three types of constraints: reduction effort (*E*), perceptual efficiency (*P*), and a limit on the overall duration of time available for speech (*D*). The definitions and violation functions of these types of constraints are summarized in Table 5 below.

Constraint	Definition	Violation function	
E	Minimize phonatory effort at the syllable level	$E = \sum_i \eta_i \sqrt{s_i}$	(square root)
P_s	Maximize perceptual clarity at the syllable level	$P_S = \sum_i e^{-\psi_i s_i}$	(inverse exponential)
P_w	Maximize perceptual clarity at the word level	$P_w = \sum_j \alpha_{w_j} e^{-\psi_i w_j}$	(inverse exponential)
D	Transmit efficiently (controls speech rate)	$D = \sum_i \delta_i s_i$	(linear)

Table 5: Constraints and cost functions used by Windmann et al. (2015), where…
s_i is the duration of the *ith* syllable,
w_j is the duration of the *ith* word,
η_i is an effort coefficient for the *ith* syllable,
ψ_i and ψ_j are perceptibility coefficients for the *ith* syllable and *jth* word,
α_{wj} is a coefficient for the strength of word prominence in the *jth* word,
and δ_i is a speech rate coefficient for the *ith* syllable.

The first constraint, *E*, which regulates phonatory effort, prefers syllable durations which are shorter, on the assumption that, over certain length, longer productions of syllables are more effortful. The authors, for abstraction, ignore factors related to articulatory effort, which they presume are only relevant for "the lower end of the temporal scale", and can be ignored when considering only syllable durations which are long enough to avoid articulatory difficulties.[16] Somewhat arbitrarily, *E* assigns violations proportional to the square root of the duration of the syllable. The violations are scaled by a coefficient η_i which allows the constraint to penalize different syllables at a different rate, presumably depending on their articulatory properties.

The second type of constraint, *P*, which regulates perceptual clarity, applies both at the level of syllables and at the level of words. It always prefers longer syllables and longer words, on the assumption that longer utterances contain more perceptual cues. However, lengthening in order to provide better cues has diminishing returns: a syllable that is, for instance, 500 ms long will likely already be very perceptible, so increasing the length further does not very much increase perception. Therefore, the authors posit that the violation function for this family of constraints is the inverse exponential function, $P_s = e^{-\psi_i s_i}$ for the syllabic version, so that it assigns one violation when duration is zero and decreasing violations as the duration increases, but with a smaller slope the greater the duration. Once again, a syllable-specific coefficient is used to make this constraint apply differently to different kinds of word and syllables, for example to differentiate between syllables with different prosodic or metrical properties. The word-level perceptual constraint, P_w, is analogous.

[16] Windmann et al.'s assumption that short syllables are always less effortful to produce is admittedly not a good one, since quicker productions will often be harder to articulate faithfully. However, the authors assume that short syllables will be phonetically reduced when this is the case, counterbalancing the increase in effort and, after this is taken into account, the effort expended on a syllable will indeed correlate with its duration.

The last constraint, D, is a constraint on the overall time of an utterance, treating time as a "shared resource." Similarly to the constraint on phonatory effort, it is violated more the longer the utterance, and contains a syllable-specific coefficient to regulate local changes in speech rate, but in this case the authors decide to make the violations it assigns a linear function of duration. The weight of this constraint will also function as the mechanism by which the grammar regulates global speech rate.

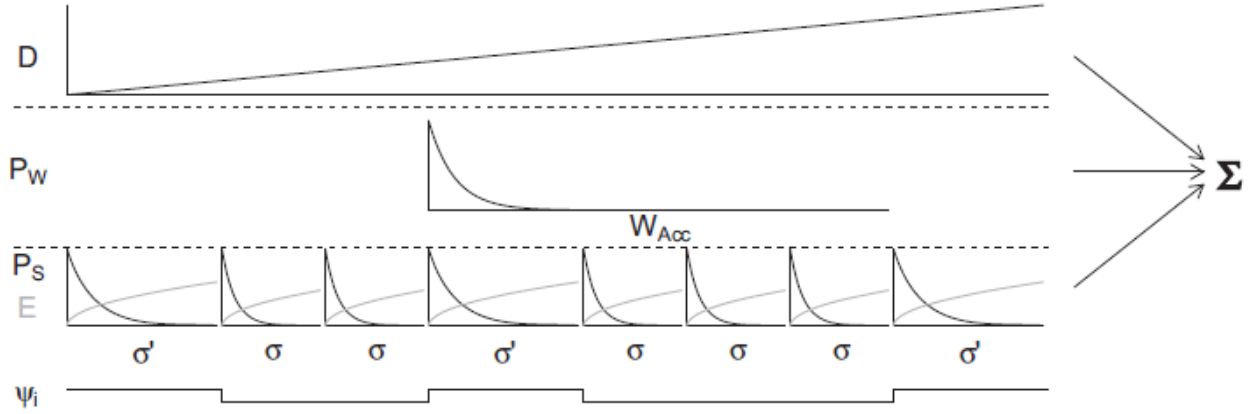

Figure 12: Figure 2 from Windman et al., 2015 (p. 82). Model architecture. Cost functions D (utterance level), P_W (word prominence; only shown for accented word W_{ACC}, as parameter x_{wj} is set to 0 elsewhere) and E/P_S (syllable level; σ; apostrophe denotes stresses) as well as stress parameter ψ_i (other parameters assumed to be constant) are plotted as a function of respective constituent durations for a hypothetical SUUSUUUS sequence. The y-axes show the costs as a function of duration (x-axis).

The overall cost of an utterance containing one or more words or syllables is the weighted sum of the violations of the articulatory, perceptual, and speech rate / efficiency constraints, scaled by how important each one is to the grammar (essentially its weight), the equation for which is given below in (12):

(15) $C = \alpha_E E + \alpha_P P + \alpha_D D$ (p. 79)

For some input with a number of syllables, given some particular values for the various coefficients, the cost is a function of the durations of each of the syllables. The set of syllable

durations which minimize the cost function is taken to be the winner. To get a sense for how this composite cost function, the sum of several very different functions, looks, consider the simplest possible case: an utterance consisting of just a single syllable. In this case, the cost function is merely a function of *s*, the duration of the syllable (13).

(16) $C = \alpha_E \eta \sqrt{s} + \alpha_P e^{-\psi s} + \alpha_D \delta s$ (excluding the P_w term for simplicity)

Figure 13 plots two versions of this function, using two values for ψ_i (in this case the values for stressed and for stressless syllables) demonstrating how changing this coefficient will result in a longer or shorter optimal syllable length.

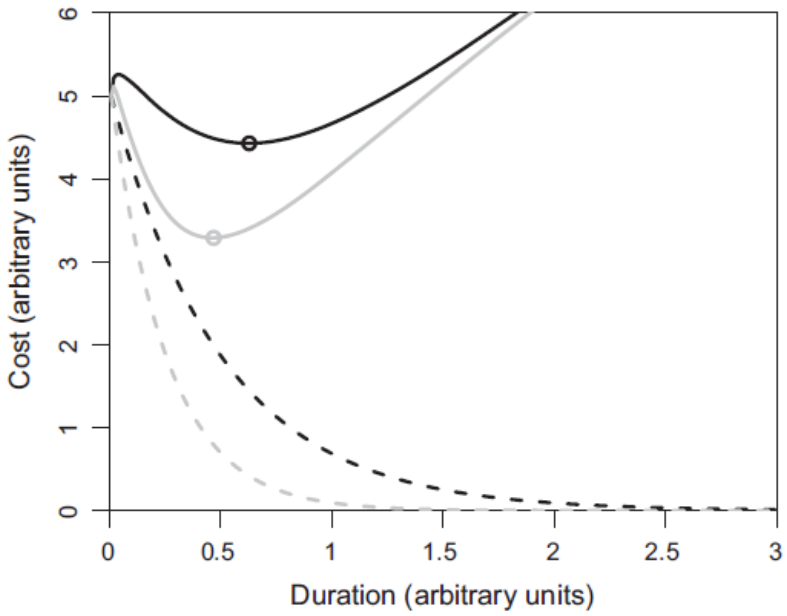

Figure 13: Figure 8 from Windmann et al., 2015 (p. 83): "Solid lines: cost function *C* (excluding P_W) for stressed (black) and unstressed (gray) syllable with above parameter settings, with circles marking optimal durations. Dashed lines: partial cost functions P_S for stressed (black) and unstressed (gray) syllable with above parameter settings."

Note that this function has two local minima: one at zero, and another which varies depending on the coefficients in the grammar. This, the authors argue, "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 ...and the absence of durations that lie in the ‘‘incompressible region” between zero and the regression intercept” (p. 83).

Windmann et al’s model also makes an interesting prediction with regards to how multiple factors influencing duration should interact. They consider the effects of accentedness on stressed syllables and unstressed syllables. As discussed in Chapter 2, stressed syllables are lengthened in accented environments to a proportionally greater degree than are unaccented syllables, which are less affected by accent, and, with the right model parameters, Windmann et al’s model predicts this to be the case.

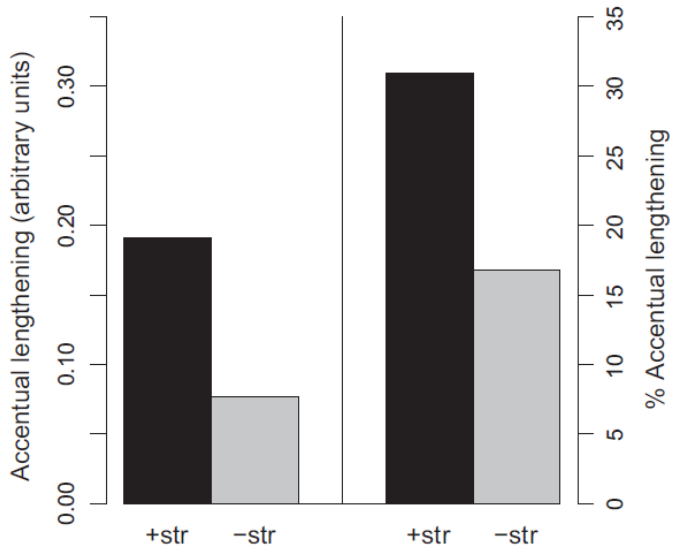

Figure 14: Figure 7 from Windmann et al., 2015 (p 85): “Absolute (left) and proportional (right) amount of accentual lengthening in stressed and unstressed syllables in the simulated utterance (bisyllabic accented word).

The authors offer the following explanation of how their model derives the asymmetry: “P_W requires lengthening of all syllables within its scope. This, in turn, allows the model to save on P_S for the individual syllables within the accented word. Yet this applies only as long as lengthening an individual syllable still makes for a sufficient reduction in P_S . Crucially, optimal duration is closer to this point for the unstressed than for the stressed syllable already without the influence

of P_W , due to the different gradients of P_S that stem from the syllable prominence parameter ψ_i. This can be observed in Fig. 8." (p 85). In other words, in an unaccented position, the location of the minimum of the composite cost function for stressless syllables is at a location where the syllable's contribution to P_S is rather shallow: lengthening a stressless syllable from this duration has diminishing returns with regards to its perceptibility. The same is less true for stressed syllables: the location of the minimum of the composite cost function in accented position is at a location where the P_S component is still quite steep: at this duration, the perceptibility of a stressed syllable would benefit a lot from lengthening. Therefore, when accented, a word level pressure to lengthen, applies to a word with stressed and stressless syllables, a synergistic response to this pressure is to lengthen the stressed syllables more than the stressless ones, since this lengthening satisfies two of the constraints (P_W and P_S) at once.

In summary, Windmann et al. outline a harmonic grammar for the duration of syllables in larger utterances which uses very different constraints than prior authors working within the phonetic HG framework. In particular, their constraints variously make use of square root, linear, and inverse exponential violation functions, and have syllable- and word-specific coefficients. Two of the constraints, *E* and *D*, have a monotonically increasing cost function, while one family of constraints *P*, has a monotonically decreasing cost function. None of the constraints have Flemming-style targets (or, stated another way, all the targets are at 0)—instead, the effect of a phonological factor (say, syllable stress) is incorporated using the coefficients, effectively adjusting constraints' weights, rather than the location of their optima. When added together, the three monotonic constraints generally sum to a composite cost function which has one local minimum at zero, and another local minimum which is not at zero or infinity, allowing the grammar (in cases where the second minimum is smaller) to predict that syllables will have a

duration which depends on the shape of the input, the constraints, and their parameters. The square root, linear, and inverse exponential violation functions were in fact admittedly chosen by the authors in part because their sum has this desirable property. Nevertheless, the eschewal of the need for constraint targets could be a desirable property for a grammar to have with respect to learnability, as will be discussed in Chapter 6.

3.4. A taxonomy

Clearly, in developing a constraint-based account for phonetic phenomena, a number of theoretical and implementational decisions need to be made. Three such decisions which are approached differently by different authors are discussed here.

Firstly, there is the question of what sort of constraint grammar framework to use. The authors just discussed have described grammars resembling Classical OT, with ranked constraints (Boersma, 2009, 2011), or Harmonic Grammar, with weighted constraints (Zsiga, 2000; Flemming, 2001; Katz, 2012; Braver, 2013; Flemming & Cho, 2017), but additional "flavors" of Optimality Theory abound: Stochastic OT (Boersma, 2003), Noisy Harmonic Grammar (Beorsma & Pater, 2008), and Maxent (Goldwater & Johnson, 2003), for example, are all variants that explicitly model variation. Orthogonally to this, each of the previously discussed versions of OT can be treated as a single parallel grammar, or a series of grammars as posited by proponents of Harmonic Serialism (McCarthy, 2000), and many other finer-grained bifurcations of the grammatical taxonomy are possible.

Secondly, there is the question of how to handle the phonetic candidate space. It can be treated as a vector space of phonetic values, or as a discretization of this space into a finite number of candidates that represent ranges or "bins" of possible phonetic values. Both approaches have their

challenges: the former necessitates a sophisticated EVAL function which calculates winners rather than selecting them from a candidate set (Flemming, 2001), while the latter arbitrarily chunks what is more naturally seen as a continuum, and may itself run into implementational problems when several phonetic values are being calculated in parallel, such that the candidate space is multidimensional, making the number of bins very large.

Thirdly, there is the issue of constructing the constraints themselves, and in particular deciding how they should assign violations. They can do so in a categorical way, akin to the constraints used in most phonological grammars, or in a continuous way. If they do so in a continuous way, it is generally a function of the degree of deviance of some phonetic measurement from some predetermined target. This target can be itself a parameter of the grammar/inventory, or it can be constant, for example set always to 0. The number of violations assigned could in principle be any function of this deviance, for example a linear function (this probably won't work well; see discussion below), a parabolic function (Flemming, 2001; Katz, 2010; Braver, 2013; Flemming & Cho, 2017), a hemiparabolic function, in which violations are only assigned when the value falls to one side of the target (this dissertation; Hayes & Schuh, MS), or a variety of heterogeneous violation functions depending on the constraint involved (Windmann et al., 2015).

Table 6 below summarizes just a part of this possible space of grammars, classifying the accounts given by some of the authors discussed. The rows represent OT formalisms, while the columns represent the constraint violation functions used. Each cell is broken into two regions representing the possibilities of discretizing the candidate space (top) for the purposes of implementing GEN and EVAL, or leaving it continuous (bottom).

	Categorical	Linear about Targets	Parabolic about Targets	Hemiparaboli c about Targets	Sqrt, Linear, Inverse Exp.
Classical OT	Boersma	2	2	2	2
	1	2	2	2	2
Harmonic Grammar	Zsiga[17]	3			
	1	3	Flemming, Katz, Braver, Flemming & Cho		Windmann et al
Stochastic OT, Noisy HG		3			
	1	3			
Maxent		3	4	Chapter 6, Hayes & Schuh	
	1	3	4	5	5

Table 6: A taxonomy of some possible formulations of phonetic constraint grammars. Where cells are divided into pairs, the top one represents an implementation with a discretized candidate space. Shaded areas denote theoretical combinations judged impossible.

It is worth noting that certain combinations, shaded in grey, either aren't workable, or are degenerate in the sense that they achieve the same effect as a much simpler grammar—these cases are discussed below. However, as can be seen, of the possible combinations only a few have been tried. I make this point to try to convince the reader that each of the proposed formalisms for

[17] While Zsiga ends up advocating for alignment constraints which use phonetic "windows", following Byrd (1996b), and Keating (1990a), the violation functions of these constraints are not discussed explicitly, and so are assumed here to be categorical (i.e. uniform within the regions inside and outside of these windows). Zsiga does not specifically mention Harmonic Grammar, but does argue for the use of weighted constraints instead of ranked ones, in response to the need to account for phonetic compromise.

phonetic grammars, Chapter 4 of this dissertation included, is best viewed as a constellation of quasi-independent theoretical decisions, and that rather than investigating the predictions and successes and failures of these accounts wholesale, it would be better to try to investigate the properties and predictions of the theoretical decisions themselves. To this end, the following paragraphs briefly discuss a few notable properties of particular regions of the taxonomy.

1) When using constraints that penalize candidates categorically based on whether the candidates phonetic value equals (or approximately equals, or falls into a window around) some phonetic value, only a finite number of candidates or categories of candidates can be distinguished from each other. For example, if the violations are binary, there are only 2^n possible violation profiles. Viewing the candidate space as continuous is not useful in these accounts, since the cost function is discrete (local regions of the candidate space will be flat with respect to how many violations they incur), so methods for algebraically minimizing the cost function are not applicable to these cases.

2) With strict ranking, the winning candidate will completely satisfy the top-ranked constraint, so only the location of the minimum of its violation function is relevant, and not the shape. This makes constraints with continuous violation functions degenerately equivalent to families of categorical constraints in which constraints barring all but one phonetic value (the optimal one) are undominated.

3) As already discussed by other authors (Flemming, 2001; Katz, 2010; Braver, 2013; Hayes & Schuh, MS), if violations are assigned as a linear function of the deviation from a phonetic target, whenever two constraints conflict with respect to some

phonetic target, the optimal candidate will be the one that exactly achieves the target of the higher ranked constraint, as shown in Figure 15.

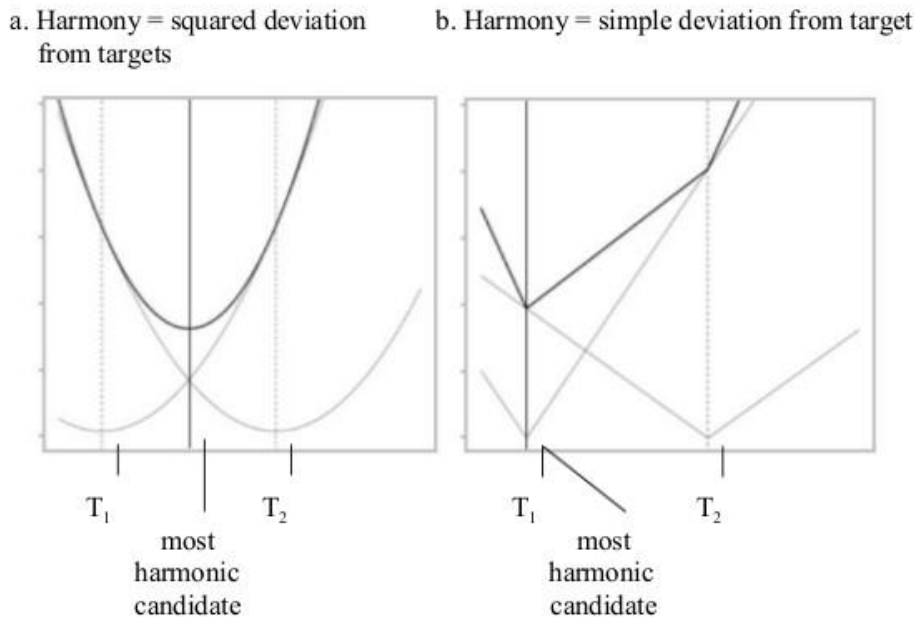

Figure 15: Taken from Hayes & Schuh (MS), p. 38; adapted from Flemming & Cho (2017): Violation functions for two constraints and their summed violations, comparing parabolic with linear violation functions. Linear violation functions fail to predict compromise between targets.

This makes these violation functions degenerately similar to categorical ones in terms of the outputs selected by harmony-maximizing grammars such as HG, contravening the primary advantage of using continuous violation functions in the first place, which is the ability to capture the empirical patterns of phonetic compromise reported throughout the literature. In OT variants which predict variation (such as maxent), however, the predictions of a model with linear violation functions are not completely degenerate: while the most probable candidate will still be the one which matches the target of the highest-weighted constraint, since variation in the output is predicted by these grammars, the other constraints will still be able to influence how often non-optimal candidates are predicted to occur.

4) If only parabolic violation functions are used in a maxent grammar, the (free-)variation of any one phonetic value in a particular phonological context is predicted to be normally distributed (section 4.1.3). This prediction of normal distributions of phonetic variables across multiple tokens is in fact an empirically testable one, and will be revisited in the discussion section of Chapter 5.

5) Calculating the probability distribution over an actually continuous space of candidates in maxent grammar (rather than simply calculating the most harmonic candidate, as in HG) is in principle possible, but runs into mathematical obstacles related to computing the normalizing constant Z that is needed to convert between $e^{harmony}$ values and probabilities, a problem that does not arise if the candidate space is discretized (section 4.1).

The following chapter extends the growing literature on phonetic constraint grammars by putting forth a proposal which investigates both a novel row of the taxonomy, moving from harmonic grammars to maxent grammars, and a novel column, moving from parabolic constraints to hemiparabolic ones, and explores some predictions about the behavior of phonetic duration that emerge from the maxent framework and from choices about constraints and constraint violation functions within this framework.

4. Maximum entropy phonetic harmonic grammars

This chapter lays out a proposal for moving from the Harmonic Grammars proposed by Flemming and others (Katz, 2010; Flemming & Cho, in press) to Maxent Grammars, also already in use by phonologists, which have the advantage of being able to learn from and model data which contains variation.[18] The feasibility of using Maxent Grammars to predict not only optimal values but probability distributions over the values of acoustic phonetic targets, such as segment duration, will be demonstrated.

The first part of this chapter discusses the application of Maxent to the realm of Phonetics, discussing the potential advantages of this approach over the use of other similar phonetic constraint grammars, as well as some of the issues related to GEN and EVAL that arise when using Maxent to select phonetic candidates as opposed to phonological ones.

The second part of this chapter lays out the constraint families for duration that will be used in this dissertation. These are similar to the constraints used by Flemming (2001), Katz (2010), and others, but differ in that they are asymmetrical, or "hemiparabolic," and that constraints can therefore specifically call for shortening or lengthening.

Finally, demonstrations of how these grammars would work in practice are provided in the form of toy examples with maxent grammar fragments, their outputs are investigated, and some of the empirical predictions of this kind of model discussed.

[18] The same approach is taken in a concurrent work, Hayes & Schuh (MS), who use Maxent grammars to model several aspects of the *rajaz* meter of Hausa, including the phonetic durations of the syllables produced. In their grammar, general linguistic constraints on duration interact with metrical constraints that enforce rhythmic consistency at several levels.

4.1. Using maxent for phonetics

Maximum Entropy (Goldwater & Johnson, 2003) is an OT framework which assigns candidates probabilities, rather than selecting a single optimal candidate. The idea is that any given candidate *x* should occur with a probability that is related to its harmony, that more harmonic candidates occur more often than less harmonic ones. This decision also allows the grammar to assign a likelihood to any particular set of observed data, and this ability is in turn the basis for a learning algorithm for these grammars.

The applicability of Maxent to the current program, modeling phonetic values with constraints, is intuitive, because phonetic values unquestionably exhibit lots of unconditioned variation across tokens. If this variation is, from a linguistic perspective, purely random, then a simpler model which predicts a single winner for each input to the grammar can be argued to be sufficient, so long as "noise" is introduced into the model. If, however, the pattern of variation is found not to be random, and in particular if the shape of the distribution of some phonetic variable is found to be linguistically meaningful, single-winner models will be at a loss to account for this without appealing to some additional mechanism. Maxent, on the other hand, explicitly models and predicts probability distributions over the candidates for each input, and in fact assigns no special status to any single optimal candidate.

Another advantage of Maxent[19] Grammars relates to their learnability. Given a set of data annotated with violation profiles, the objective function over the space of constraint weights is

[19] Though Flemming and Cho (2017) automatically learn the weights and targets of their HG grammar fragment by treating the single "winner" phonetic values output by the grammar as instead being the centers of normal distributions, such that experimental data can be assigned probabilities as a function of particular model parameters rather than simply matching or not matching the output, and the best model can be learned via log-likelihood maximization, much as in Maxent learning.

provably convex, and the weights which maximize the likelihood can therefore be easily learned by any number of optimization algorithms (Goldwater & Johnson, 2003; Hayes and Wilson, 2008).

4.1.1. Computing probabilities in maxent phonetic grammars

In order to understand the challenges unique to applying the maxent grammar formalism to phonetic variables like duration, it is first necessary to understand the processes by which maxent assigns probabilities to candidates in the more familiar case where the candidates are categorical. First, the harmony for each candidate is computed by taking the weighted sum of its violations (here assumed to be a positive number, where a larger number is less harmonic, and 0 is the most harmonic value; note that in some other accounts harmonies are treated as negative numbers). Then, for each candidate *x*, a "maxent value" equal to $e^{-harmony(x)}$ is generated. The probability of the candidate occurring is taken to be exactly proportional to this maxent score, such that a candidate with double the maxent score of some other candidate is exactly twice as likely to occur.

However, the maxent scores of the candidates do not sum to one, so even though they are proportional to these candidates' probabilities, in order to convert them into actual probabilities, they have to be normalized. To do this, a normalizing constant *Z* is computed by summing over the maxent values of all candidates, and the individual maxent values are then divided by *Z*.

/UR/	C_1 (w = 2)	C_2 (w = 5)	harmony	$e^{-harmony}$	probability
[x_1]	2		2	0.135	0.946
[x_2]		5	5	0.00674	0.047
[x_3]	2	5	7	0.000912	0.006
				Z = 0.143	

Table 7: Maxent grammar example.

In the phonetic domain, if the candidate space is taken to be a continuous vector space—a 1-dimensional space if the grammar is determining a single phonetic value, or a multidimensional one if more than one phonetic value is being determined in parallel—the situation is somewhat different. As in the categorical case, any one candidate can easily be assigned a harmony and a maxent value by assessing its violations. However, doing so for each of the infinitely many candidates in the candidate set is obviously not feasible. Furthermore, even if one were to do so, the sum of their maxent scores would be infinite, and the probability of any particular candidate would therefore necessarily be zero. This makes sense: if candidates are point objects in a vector space, they should each only be infinitesimally likely. However, *regions* of the candidate space do have probability. For example, while the probability that a vowel will turn out to be exactly 89 ms long is infinitesimal, the probability that it will be between 89 ms and 90 ms is not.

This state of affairs leads to two different approaches to computing probability distributions over the infinitely many candidates, which are described in the sections that follow.

4.1.2. Computing probability with a discretized candidate set

The first approach is to simply discretize the candidate space, à la Boersma (2009). If the vector space, be it one- or multi-dimensional, is broken into bins, the set of (plausible) candidates becomes finite. For example, if the duration of one segment is being modeled, the candidate set can be approximated by a list of the durational regions 0-5ms, 5-10ms, 10-15ms, etc. These bins could be larger or smaller, but in any case, there will be a finite number of them, assuming no

negative durations are possible,[20] and that there is also some upper bound on phonetic duration.[21] The process of computing probabilities in this case is just the same as in the phonological case.

To illustrate, consider a grammar which has a single constraint, with a weight $w = 1$. Following Flemming and others, let's set the cost of violating this constraint (and therefore the total cost in this grammar) to be the square of distance between the duration x of the candidate (expressed in centiseconds, for ease of illustration), and a target duration value, $t = 7.5$ cs. If the candidate space is discretized with a resolution of 2 cs, and capping duration at 20 cs, the following tableau contains the complete list of candidates, along with their violations, harmony, maxent score, and probabilities.

[20] Durations of exactly 0 are plausible if some cases of deletion are taken to be phonetic processes rather than phonological ones (Windman et al., 2015), for example where some process of phonetic reduction applies to a variable degree and can result in apparent deletion in the most extreme cases. A hypothetical negative segment duration could perhaps be thought of as an even more extreme case, where a drive for phonetic syncope is so strong that it is also responsible for gestural overlap between the (now-adjacent) surrounding sounds, but this is far-fetched.

[21] The Guinness World Record for longest continuous vocal note is, at the time of writing, an [u] produced in 2016 by a Turkish man named Alpaslan Durmuş for 1 minute and 52 seconds, so it's probably safe to exclude durations exceeding 2 minutes from the candidate set. Alternatively, since any descriptively adequate grammar should already penalize very long durations, there will necessarily be some duration above which the grammar assigns less than 0.000000001% of the probability mass, such that excluding candidates longer than this duration will have a negligible effect on the predicted probability distribution.

[x]	DUR[x] w = 1, t = 7 cs	harmony	$e^{-harmony}$	probability
0-1 cs	49	49	0.0000	0.0000
1-2 cs	36	36	0.0000	0.0000
2-3 cs	25	25	0.0000	0.0000
3-4 cs	16	16	0.0000	0.0000
4-5 cs	9	9	0.0001	0.0001
5-6 cs	4	4	0.0183	0.0103
6-7 cs	1	1	0.3679	0.2075
7-8 cs	0	0	1.0000	0.5641
8-9 cs	1	1	0.3679	0.2075
9-10 cs	4	4	0.0183	0.0103
10-11 cs	9	9	0.0001	0.0001
11-12 cs	16	16	0.0000	0.0000
12-13 cs	25	25	0.0000	0.0000
13-14 cs	36	36	0.0000	0.0000
14-15 cs	49	49	0.0000	0.0000
$Z = 1.7726$				

Table 8: Violations maxent values, and probabilities for a single-grammar constraint with a discretized candidate set.

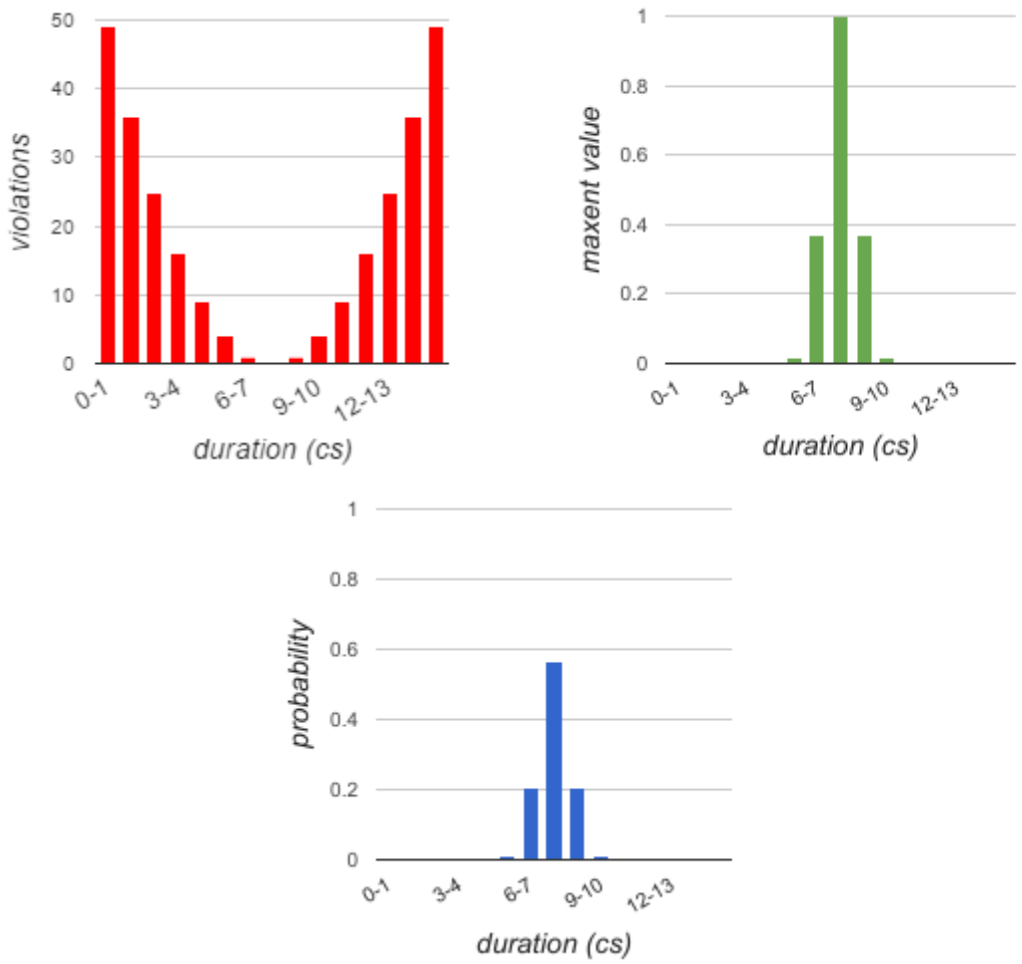

Figure 16: Violations, maxent values, and probabilities as a function of duration for a single-constraint maxent grammar with a discretized candidate set

Note that, using this discretized approach, we are in effect creating a model of what will appear in a *histogram* of the phonetic values that will be observed for a given input across multiple tokens.

4.1.3. Computing probability density with a continuous candidate set

The second approach is to leave the candidate space continuous, treating each candidate as a point object. The violations assigned are already a continuous function of candidate duration, as are the maxent values, so there is no need to compute these on a candidate-by-candidate basis.

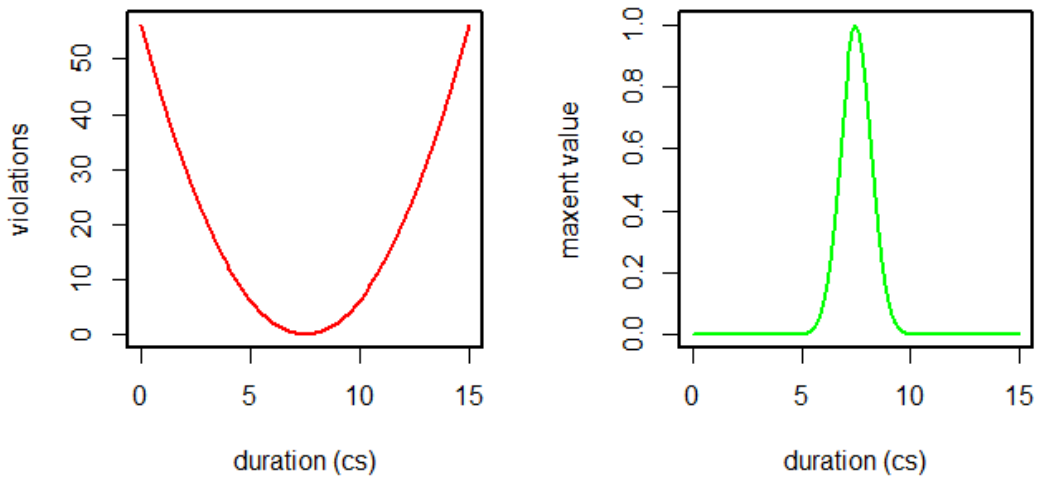

Figure 17: Violations and maxent values as a function for duration for a single-constraint maxent grammar with a continuous candidate set.

However, since the probability of each possible phonetic value is infinitesimal, a change needs to be made to our conception of the relationship between maxent values and candidate probability. In particular, the conceptual change is as follows: instead of taking a candidate's maxent value to be proportional to its probability, we must instead take it to be proportional to that candidate's *probability density*.

For readers unfamiliar with probability density, it may help to momentarily consider what is meant by density in the physical realm. Imagine for a moment that you are a Greek philosopher unaware of (or perhaps just opposed to) the idea of atomism, and believe simple substances, like water or lead, to be undifferentiated masses. Being familiar with the work of Zeno of Elea, however, you realize that since volumes of these substances could theoretically be halved, and then halved again, forever, that there must be an infinite number of parts to even a small object. Since the weight of an object is distributed throughout it, each of these parts, you reason, must have weight. However, since there are infinitely many of them, each must have, on average, no weight at all. Further complicating matters, similar volumes of water and of lead have very

different weights, despite both being composed of infinitely many weightless subparts. While one solution to this paradox would be to wait around for Isaac Newton to invent calculus, a stop-gap measure would be to conceive of weight as coming not from the many individual points in a substance, but collections of them, where the relationship between a volume of points and its weight is the substance's density. Note that density need not be the same throughout an object: different parts of the object could have different densities, and the density could even vary continuously throughout the object (though in this last case computing the weight probably would require waiting for Isaac Newton, or at least the use of a scale).

The same idea is applicable to the notion of probability distributions over vector spaces of phonetic values. Each point in phonetic space, rather than having a probability, will have a probability density: a ratio between the volume of the infinitesimal region around that candidate in phonetic space and the probability of that region, and it is this value which is computed by the grammar for any particular candidate. As a concrete example, take our single-constraint grammar for duration. Figure 18 graphs the probability density in Hz[22] of the candidates.

[22] The reader may be wondering why hertz have suddenly appeared in a dissertation on duration. The reason is that probability density is probability over some phonetic space, and, since the probability component is unitless, it will have units that are the inverse of the units of the phonetic space in question, which in this case are units of time. 1 Hz can be thought of as equivalent to the probability density of a candidate duration in a continuous uniform distribution 1 second wide.

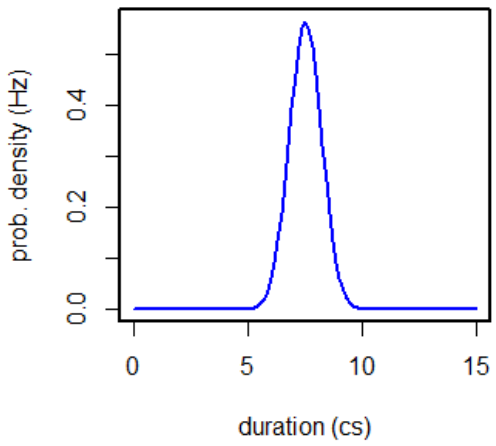

Figure 18: Predicted distribution for a single-constraint maxent grammar with a continuous candidate set.

Probability, a property of *ranges* of candidate durations, corresponds to the area under the relevant portion of the maxent value curve. The total area under the curve, therefore, must equal 1. Since we take maxent values to be proportional to probability density, computing probability density simply involves dividing the maxent values by a normalizing constant *Z*, which must be the total area under the maxent value curve. Finding *Z* in these grammars therefore involves taking the definite integral of the maxent value function, itself the exponential function of the harmony function of the candidates.

(17) $h(x) = (x - t)^2$

(18) $P(x) = e^{-h(x)} = e^{-w(x-t)^2}$ where P(x) is the maxent value of x

(19) $Z = \sum_x P(x) = \int_{\infty}^{\infty} e^{-w(x-t)^2} = \sqrt{\frac{\pi}{w}}$

(20) $pd(x) = \frac{P(x)}{Z} = \sqrt{\frac{w}{\pi}} e^{-w(x-t)^2}$

...where $P(x)$ is the maxent value of the candidate with duration x, and $pd(x)$ is its probability density. This is exactly the equation for the normal (or Gaussian) distribution, with the mean of the distribution $\mu = t$, and the variance of the distribution $\sigma^2 = 2/w$.

$$(21) \quad f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A famous fact about the function $e^{-(x-\mu)^2}$ is that its definite integral, taken from negative infinity to infinity, is $\sqrt{\pi}$. This is notable because the integral of this function does not have a closed-form solution in the general case (its integral is rather opaquely named the "error function"), so that if the definite integral of such a function is taken over other ranges of values, like 0 to infinity, or 0 to 20, it can only be approximated.

Because sums of violations due to constraints with parabolic violation functions like the ones proposed by Flemming (2001) will themselves be parabolas, Maxent phonetic grammars consisting of only these sorts of constraints will always output, for any particular phonetic variable, a predicted distribution which is Gaussian in nature.

4.1.4. Excursus on the feasibility of continuous candidate sets

While this method of computing Z appears to work for our very simple grammar, the feasibility of implementing this step more generally depends on the violation functions that are chosen for the constraints, and especially on the number of phonetic values being computed, since additional phonetic variables in the candidate will result in a multi-dimensional candidate space, making the space under the maxent curve a volume rather than an area. In this case a *multiple integral* would need to be taken to compute Z.

As a concrete example, consider the grammar from Flemming (2001) governing CV coarticulation. The two phonetic values that are output by the grammar are the F2(C), the F2 locus of the consonant, and F2(V), the second formant of the vowel, and these values are determined by the interactions of three constraints: ID(V), ID(C), and MINEFFORT. The overall cost (i.e. harmony) in this grammar is given by (15).

(22) $C = w_c(F2(C) - L)^2 + w_v(F2(V) - T)^2 + w_e(F2(C) - F2(V))^2$ p(20)

For visual clarity and abstractness, let us replace the two phonetic values governed by this grammar with *x* and *y*, the constraint weights with w_1, w_2 and w_3, and the targets and loci with t_1 and t_2. From the cost (i.e. harmony) function, we can calculate the maxent value function *P*(*x*,y). This function is plotted in Figure 19.

(23) $h(x, y) = w_1(x - t_1)^2 + w_2(y - t_2)^2 + w_3(x - y)^2$

(24) $P(x, y) = exp(-w_1(x - t_1)^2 - w_2(y - t_2)^2 - w_3(x - y)^2)$

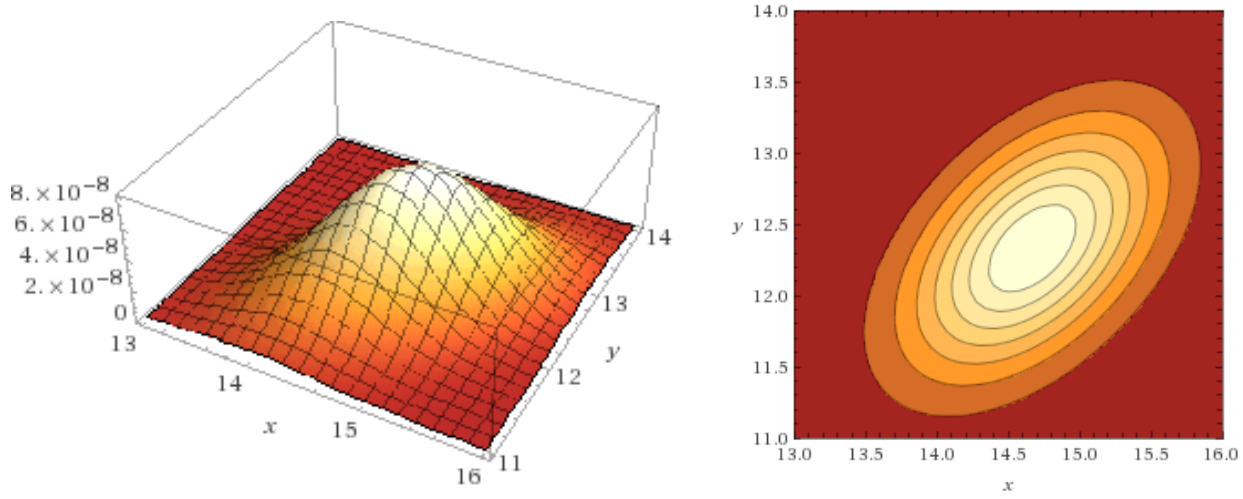

Figure 19: Maxent value as a function of F2(C) and F2(V) (*x* and *y*, respectively) in a maxent grammar with ID(V), ID(C), and MINEFFORT, with the consonant F2 locus set to 1700 Hz, the vowel F2 target set to 1000 Hz, and all weights set to 1, following Flemming (2001). Durations are represented in hectohertz (10^2 Hz), rather than hertz, for visual clarity.

Computing Z in a more complex case like these will, since Z is now a volume, necessitate integrating first over one of the phonetic variables, and then over the other. However, the result of integration by just one of the variables is not itself always integrable (depending on the constraints' functions, and the range of durations considered). This presents a significant mathematical roadblock. While a way of reliably computing Z for maxent curves in arbitrary-dimensions may exist, after much effort, I have not been able to find one.

While computing Z in order to normalize the height of the maxent value curve may seem like a formality, it is essential for maxent learning: fitting the parameters of such a grammar to data will necessarily involve computing Z many times for many tableaux with different parameter values. Since one of the goals of this dissertation is to algorithmically learn the parameters of maxent phonetic grammars, the continuous candidate space approach was ultimately abandoned in favor of discretizing the candidate space into "bins" (the first approach discussed in this section) and the implementation of the learning algorithm provided in Chapter 6 reflects this decision. To readers skeptical of the discretized candidate space approach (which admittedly is less mathematically elegant), I point out that, as the bins are made smaller, the predictions of the discretized approach converge with those of the continuous approach, and so the former can be viewed as an arbitrarily good approximation of the latter with fewer implementational barriers.

4.1.5. Constraint violations and phonetic distributions

The proposals for phonetic constraint grammars discussed in the previous chapter all have the property that they predict a single winner. For such models, the variation in phonetic values across tokens generated by the same grammar must therefore be explained via another mechanism, such as noise. For maxent phonetic grammars, however, phonetic token variation is a direct result of

the grammar. The exact shape of the distribution of any particular phonetic value depends on the constraints that regulate that value, and in particular on their violation functions.

If all the constraints assign violations in a parabolic fashion, the distribution of any given phonetic variable will be a normal (or Gaussian) distribution. This is because the harmony of a candidate will be the sum of a number of parabolic violation functions, and the sum of several parabolas is itself a parabola. Thus, limiting violation functions to be of this form in a maxent grammar makes a strong empirical prediction: that random variation in phonetics, holding phonological factors constant, should be normally distributed.

The use of other violation functions, for example the inverse exponential, square root, and linear functions employed by Windmann et al. (2015), would under a maxent account predict very different durational distributions: ones related to the inverse exponentials of these violation functions and their sum. For example, the harmony function given by their model for a single stressed syllable in isolation (using the parameters from p. 84) is given in (22), and the corresponding maxent value function in (23), both also plotted in Figure 20. Note that since this harmony function has two local minima—one at zero and one at some positive duration—the predicted distribution is bimodal, predicting deletion or near deletion in some tokens and a range of positive values in other cases, to the exclusion of durations in between.

(25) $h(x) = 3\sqrt{x} + 5e^{-2x} + x$

(26) $P(x) = e^{-h} = exp(-3\sqrt{x} - 5e^{-2x} - x)$

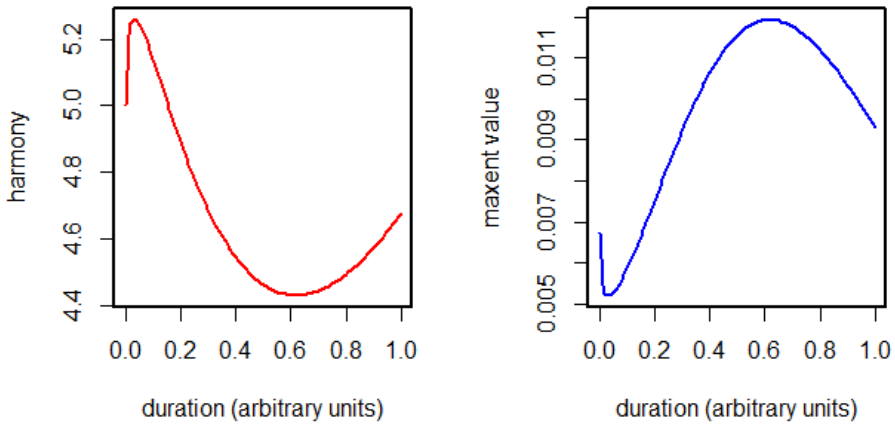

Figure 20: Harmony and maxent value functions in a hypothetical maxent grammar with the constraints proposed by Windmann et al. (2015).

While the examples above demonstrate how the shape of predicted probability distributions is related to the choice of constraint violation functions, the maxent framework also makes certain predictions which do not depend on this choice. One such prediction involves the consequences of adjusting constraint weights in these grammars. In the simpler phonetic HG framework, where only one winner is predicted, the weights of phonetic constraints essentially determine how much the phonetic variables they constrain should adjust in response to pressure from competing constraints: phonetic variables governed by constraints with large weights will stay close to the targets of those constraints in a variety of phonological contexts, while those subject to constraints with small weights will show more accommodation to their phonetic or phonological context (i.e. to competing constraints). The weight parameter in these grammars is therefore empirically justified in so far as there are some categories of sounds which have realizations that are more consistent across contexts, while others show more variation.

In maxent phonetic grammars, however, the weights of constraints do double duty: in addition to governing the amount of accommodation that will occur in response to different conditioning environments, the weight of a constraint will also determine the amount of "random" variation

seen in the durations of the tokens resulting from a single input. This means that the maxent framework has a mechanism for modeling situations in which the realization of some sound or class of sounds, even in a single context, shows less variance than some other sound or class of sounds.

4.1.5.1. The Consistent Variation Hypothesis

Because the weights of constraints do double duty in this way, the maxent phonetics framework in fact makes a strong general empirical prediction: the amount of variation that phonetic realization some class of sounds shows across phonological contexts should be directly related to the amount of "random" variance it shows within any one phonological context. In other words, the sounds whose realizations have very high standard deviations, even controlling for linguistic context, should be just the sounds that adjust their realizations more readily in response to external phonological conditions. Conversely, sounds which exhibit less variance should also be less responsive. This prediction is termed the "Consistent Variation Hypothesis."

The Consistent Variation Hypothesis

When one category of sounds (or larger prosodic constituents), defined either by its phonological properties or by the context in which it occurs, shows more random, unconditioned variance in some phonetic variable than some other comparable category, it should also show more phonologically conditioned variation than that other category, appearing more susceptible to orthogonal phonological factors.

No such prediction about the correlation between variance and conditioned variation is made by existing models of duration (including phonetic HG), which predict duration means, or a single winner, but make no special predictions about the shapes of probability distributions.

Empirical evidence suggesting that this hypothesis is correct, in the form of patterned variation in the shape of durational distributions, will be presented in Chapter 5.

4.2. STRETCH and SQUEEZE: hemiparabolic constraints

The constraint DURATION, used by Flemming and others, assigns violations in a parabolic fashion, based on distance from a duration target. It is formulated such that any given constraint is violated both by candidates which undershoot the target and those that overshoot it, and to an equal degree. When the targets in question are articulatory, or are acoustic targets which correspond well to articulatory targets, such as formant values, this is a plausible (though certainly unproven) assumption. With duration targets, it's less clear that this symmetry is a good a prior assumption. Most processes relating to duration are currently thought to be directional, having either a lengthening or a shortening effect compared to what is assumed to be the default duration for a given unit. With symmetrical phonetic constraints, this would not be so. For example, the constraint that implemented phrase-final lengthening would penalize candidates under a certain duration, but also candidates over that duration. While overlong candidates should indeed violate some constraint in the grammar, it is perhaps strange to think that they violate a constraint which motivates lengthening.

However, the DURATION constraint family can be generalized slightly by splitting it into two families of constraints called STRETCH and SQUEEZE, which are explored in this chapter. Like DURATION, constraints in each of these families have two parameters: a structural description that must be met in order for the constraint to penalize the phonetic duration of some part of the candidate (for example, "high vowel", or "phrase-final syllable"), and a durational target. Also like DURATION, these constraints assign violations proportional to the square of the

distance between the duration of the relevant part of the candidate and its durational target, except that STRETCH and SQUEEZE are asymmetrical: STRETCH penalizes durations only if they are shorter than the target duration, assigning 0 violations to candidates longer than this target, and the reverse is true for SQUEEZE. Each constraint, therefore, has a "hemi- parabolic" violation function.[23]

	Definition	Violation Function
STRETCH[x, t]	The duration of a part of the speech signal matching the structural description *x* should have at least duration *t*.	$(d - t)^2$ if $d < t$ 0 otherwise
SQUEEZE[x, t]	The duration of a part of the speech signal matching the structural description *x* should have at most duration *t*.	$(d - t)^2$ if $d > t$ 0 otherwise

Table 9: Definitions of the durational constraints STRETCH and SQUEEZE.

[23] Since STRETCH is a constraint which is violated by segments shorter than a certain duration, it could equivalently be named *SHORT, and SQUEEZE as *LONG; the reader is invited to conceive of these constraints in whichever way is more intuitive.

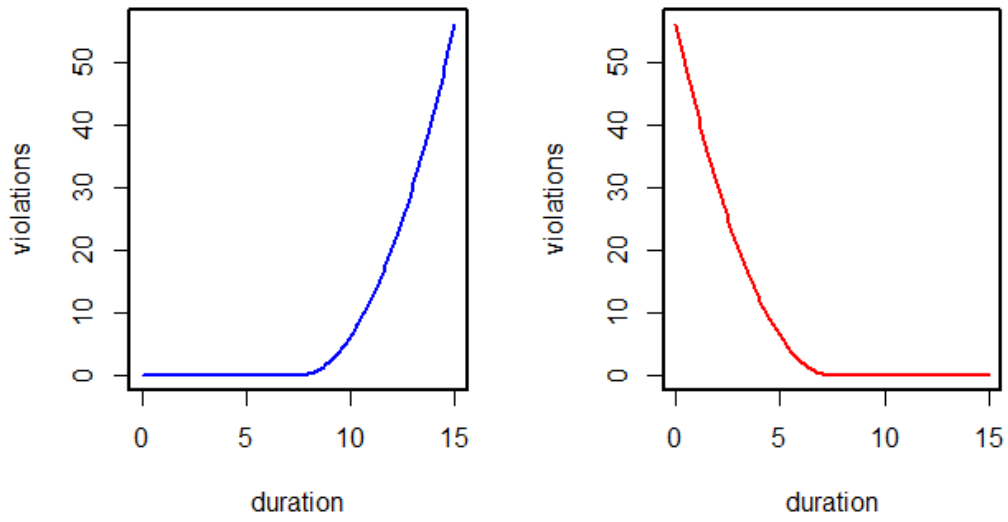

Figure 21: Hemiparabolic violation functions for STRETCH (left) and SQUEEZE (right), with weights set to 1 and targets set to 7.5.

Note any given DURATION[x,t] constraint can be mimicked by a pair of STRETCH and SQUEEZE constraints with the same structural description, target, and weight. The grammars possible using these constraint families, therefore, are a superset of the grammars that would be possible using only parabolic duration constraints.

This asymmetry allows the grammar to account for a number of hypothetical possibilities. For example, one could imagine a situation in which the duration of some sound or prosodic constituent occurs in an asymmetrical distribution, where there are plenty of tokens with duration above the most frequently observed value, and some tokens with durations that are much larger, but almost no tokens with durations below this value, or the converse, where durations longer than the optimal one seem to be dispreferred. This situation can easily be generated by a pair of STRETCH and SQUEEZE constraints with the same structural descriptions and durational targets, but with different weights, as in Figure 22.

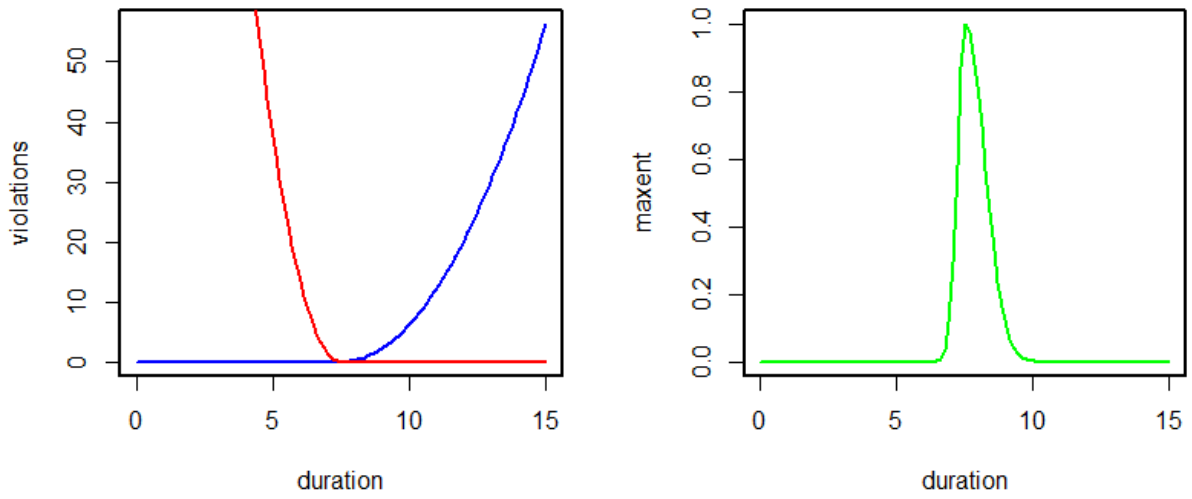

Figure 22: Violations incurred by a STRETCH constraint with a high weight (red) and a SQUEEZE constraint with a low weight (blue), and the maxent values for candidates subject to both constraints (green). The resulting distribution has a positive kurtosis / skew.

Because the overall cost function, the sum of STRETCH and SQUEEZE, is in this case steeper on one side of the optimal value, so too will be the probability density function, which will resemble the left and right halves of two Gaussians with different variance but the same mean, as seen in Figure 22. In other words, the predicted distribution will be skewed.

Divvying up the work of constraining duration into STRETCH and SQUEEZE also allows the straightforward implementation of several ideas about phonetic targets that are already found in the literature. For example, in the phonetic window model proposed by Keating (1990a) and others, a phonetic value like duration can vary freely (at least in so far as one constraint is concerned) within some range, but should not be outside this range. A “soft” window can be constructed easily using a STRETCH and a SQUEEZE constraint with the same structural description but different targets and relatively high weights. The location and size of the window will correspond to the targets selected, and the egregiousness of falling outside this window on either side will correspond to the constraint weights. Within the window, the optimal phonetic value (or distribution of phonetic values) might then be determined by unrelated constraints.

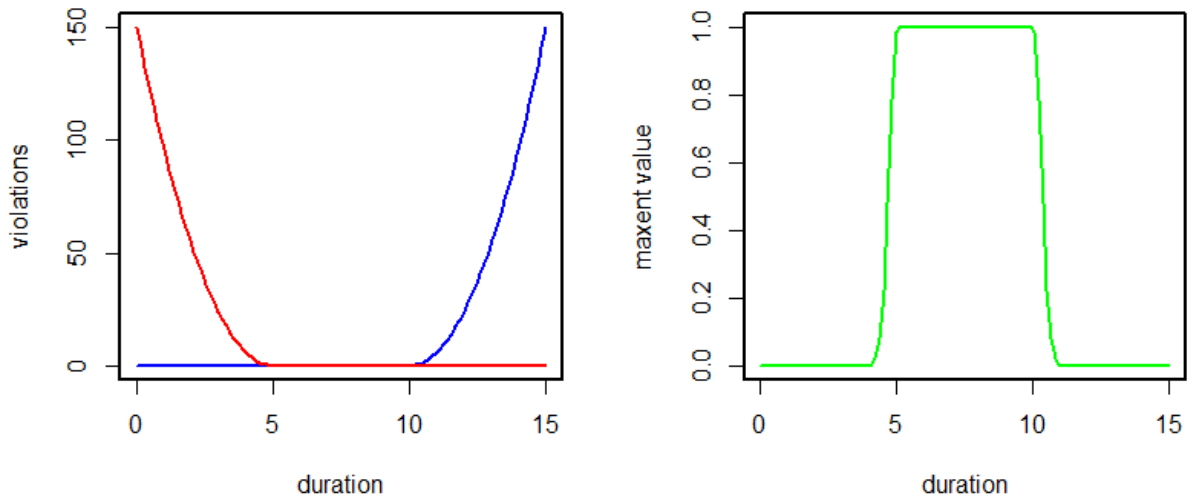

Figure 23: A "soft" window model of duration, composed of a STRETCH constraint with a shorter target (red) and a SQUEEZE constraint with a longer target (blue). Maxent values for candidates subject to both constraints are shown in green: the predicted distribution (at least without any other constraints) is uniform variation within a duration range, with a small number of outliers, depending on the constraint weights.

"Soft" maximum compressibility effects can be similarly modeled, since they are analogous to phonetic windows that open on one side. All that is needed to capture maximum compressibility is a STRETCH constraint with a short target but a very high weight: this will effectively enforce a minimum duration without imposing any preferences on the duration if it is above this minimum, allowing variation in the region where there is no danger of over-compression to be governed independently by separate STRETCH and SQUEEZE constraints. Note that while these constraints seem to create "walls," which can be arbitrarily steep, their violation functions nevertheless still have slopes which are locally continuous, a property which will be helpful when it comes to learning grammars with these constraints.

4.3. The unbounded duration pathology

In order for a particular grammar composed of STRETCH and SQUEEZE constraints to function properly, the structural descriptions of the constraints must be such that adding duration

to any part of the output must violate at least one SQUEEZE constraint. If this is not the case—that is, if there is some segment in the output which neither violates a SQUEEZE constraint nor is part of some larger prosodic constituent that does so, then all durations above the targets of the STRETCH constraints violated by this segment (if any) will be equally harmonic (and thus equally likely to occur). Since duration could then be arbitrarily long, the area under maxent value curve will not be defined, and it won't correspond to a probability distribution.

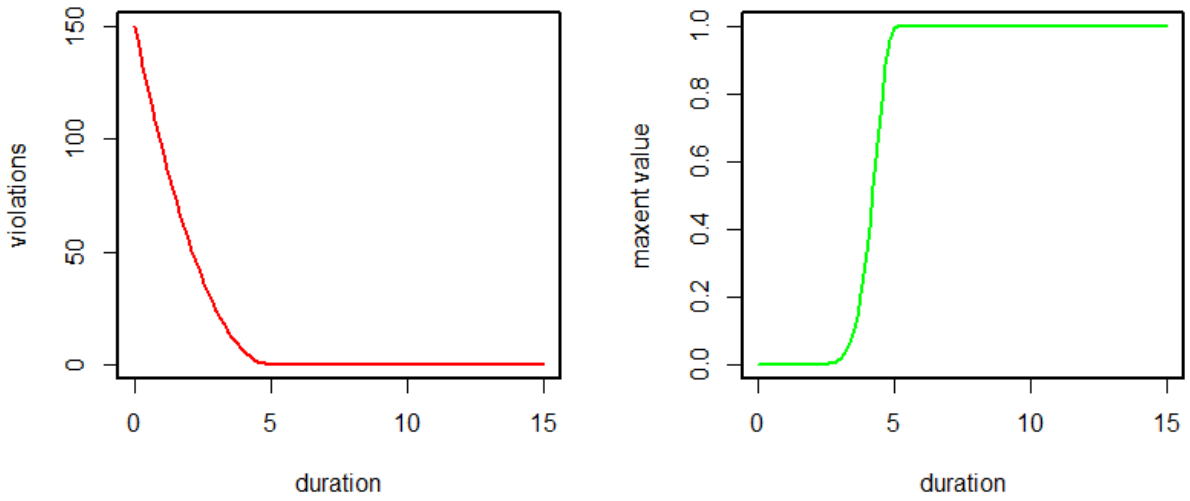

Figure 24: A pathological grammar with only a STRETCH constraint. Since the area under the maxent value curve is not defined, the probability distribution cannot be calculated.

This is in fact a special case of a pathology described by Daland (2015) that occurs in maxent grammars generally whenever a candidate space is under-constrained—in particular, when some infinite subset of the candidates is equally harmonic. The solution proposed by Daland, who addresses this pathology as it arises in maxent grammars for phonotactic well-formedness, is to require a non-zero weight for a general structure-penalizing constraint, *STRUCT, which is violated by outputs in proportion to the amount of phonological material they contain.

In the phonetic grammars considered, along the same lines, it is sufficient to posit that the grammars must have a either a pervasive SQUEEZE[S] which applies to all segments, or else a

constraint like SQUEEZE[σ] or SQUEEZE[Ft] which applies to all material at some higher level of the prosodic hierarchy, and which has a non-zero weight. Since all parts of the durations of all candidates will contribute to violations of these structure penalizing constraints, the pathology will be avoided.

The independent need for a phonetic constraint prohibiting overly long productions seems obvious upon reflection: regardless of the specific grammatical properties of sounds being produced, producing them for an extremely long time should always be marked, since doing so would result, at best, in communicative inefficiency resulting from an undue expenditure of articulatory effort, and, at worst, in speaker death by suffocation. A strong bias toward obeying something functionally equivalent to a universal SQUEEZE constraint is therefore more or less mandated by Darwinian natural selection.

4.4. Illustration of a maxent grammar for duration

The following section has two purposes. The first is to further illustrate how durational maxent grammars with STRETCH and SQUEEZE can be used to model the duration and timing aspect of the speech signal by investigating cases more complex than the ones in the previous section, where more than one input to the grammar is considered. The second, however, is to explore what kinds of patterns in the realization of phonetic duration might be predicted to occur as necessary consequences of the maxent formalism and the constraint families hypothesized, and to begin to probe what can be thought of as the phonetic equivalent to a "factorial typology" for duration.

In chapter 2, one of the central points was the importance of understanding how multiple factors interact with each other to influence duration, and how little we know about this subject empirically. To get a sense of how multiple factors would interact in maxent grammars with

STRETCH and SQUEEZE, let us use this formalism to model the data discussed by Klatt (1973b) in his paper specifically investigating effect interaction. In Klatt's experiment, vowel duration is influenced by two independent variables: the [+/-voice] feature of the following consonant, and whether it occurs in a monosyllable or in the first syllable of a trochaic word.[24] Klatt takes the unmarked case to be the phonetically longest one—vowels in a word-final syllables followed by voiced consonants—and the two effects at play to be monosyllabic (or non-word final) shortening, and pre-voiceless shortening.

	1σ	2σ
/__[-voice]	132 ms	103 ms
/__[+voice]	198 ms	131 ms

mean observed durations

	1σ	2σ
/__[-voice]	(132 ms)	86 ms
/__[+voice]	(198 ms)	(131 ms)

durations predicted by Klatt's original log-linear model

Table 10: Observed and predicted vowel durations, by context, in Klatt, 1973b.

In his paper, Klatt makes the case that, with the right grammatical formalism, the duration of the vowel in the upper right cell, [2σ -v], where both effects are present, should be predictable from observing the base duration [1σ +v], and the duration that results when each of the effects apply independently, [2σ +v] and [1σ -v], since this should be enough to estimate the size of the effects, and the model should do the rest. Klatt's original model, treating each effect as a multiplier on an inherent duration, turned out to significantly underpredict the length of the shortest category,

[24] In Klatt's data, since the two categories of words being tested are monosyllables and trochees, to the exclusion of iambs, word-finality of the syllable is conflated with polysyllabicity, so it is unclear to what extent each is at play here. Umeda (1975) finds that the final syllables of iambic words have similar durational characteristics to monosyllables, so the difference between the categories in Klatt's data may be mostly due to word-final lengthening / word-medial shortening, rather than a factor of the number of syllables in the word.

over-predicting the amount of shortening that should occur when both effects were present, and leading him to add additional parameters to his duration equation in subsequent models related to the hypothesized minimum durations of vowels.

Let us here undertake a similar endeavor using the formalism of maxent phonetic grammars. The goal will be to employ a number of STRETCH and SQUEEZE constraints, constructing the simplest possible constraint grammar of this sort which can account for three cases where none or just one of the two shortening effects is present, and, as Klatt did, investigate the predictions of the grammar with respect to the fourth case where they interact. A constraint grammar of just this sort is proposed in Table 11.

Constraint	Target	Weight
STRETCH[V]	t = 0.198 s	w = 155
SQUEEZE[V]	t = 0.198 s	w = 155
SQUEEZE[V/__σ]	t = 0.052 s	w = 155
SQUEEZE[V/__[-voice]]	t = 0.073 s	w = 155

Table 11: Constraints for a grammar governing the interaction between two durational effects.

To explain how this grammar was constructed, first consider the word type that Klatt presumes to be the base case, [1σ +v] (i.e. the vowels in monosyllabic words before voiced consonants). If we assume that this case is the least marked, it should also be subject to the fewest constraints. However, for the reasons discussed in this chapter, even the duration of a completely unmarked segment must be subject to general constraints, or the distribution of its duration will be undefined. To this end, let us posit the presence of constraints which govern the duration of vowels generally, without respect to their context, namely STRETCH[V] and SQUEEZE[V]. Since these are the only constraints that apply to vowels in the unmarked [1σ +v] category, in order to get the mean

duration of this category to match Klatt's data, the targets for these constraints must be set to 198 ms. Since we don't know the distribution of Klatt's data points, let us make the assumption that they are normally distributed, and that therefore the two constraints should also share the same weight, together effectively forming a single parabolic duration constraint.

Picking a reasonable number for the weight of these constraints requires making a guess as to the standard deviation of the duration of the vowel for this category, since constraint weight directly governs variability in these grammars. While the standard deviation for Klatt's data is not reported, an experiment in Chapter 5 tests the duration of a variety of vowels in a large variety of segmental and prosodic contexts, including contexts in which the vowel was in a monosyllable followed by a voiced obstruent, and was phrase-medial, and was pitch-accented, closely matching the [1σ +v] condition in Klatt's experiment. The mean duration for this subset of the data was 182 ms (remarkably similar to Klatt's result) and the standard deviation was 57 ms, so we can reasonably choose 57 ms as the standard deviation that our constraint weights should derive. If the data are expressed in seconds,[25] the weight w for STRETCH[V] and SQUEEZE[V] that corresponds to this standard deviation is $w = 155$. For reasons of simplicity, and to further constrain the parameters of the model, all the constraints in the grammar are given this same weight.

To model the two independent shortening effects that apply when a vowel occurs in a two-syllable word (i.e. is in a non-word-final syllable) or occurs before a voiceless obstruent, two additional constraints SQUEEZE[V/__σ] and SQUEEZE[V/__[-voice]] are needed. Since the [1σ

[25] The weights learned by a grammar depends on the units in which duration is measured. This is simply because the steepness of the parabolic violation function similarly depends on the units of the x-axis, so a constraint which assigns some number of violations to a candidate deviating from its target by one second will need to be made 1,000,000 ($1{,}000^2$) times weaker to assign the same violations if this deviation is instead expressed as 1,000 ms. However, the ratios of the constraint weights, and therefore the general shape of the grammar, is (reassuringly) unaffected by the choice of units.

-v] and [2σ +v] cases turned out in Klatt's experiment to have very similar average durations (132 ms and 131 ms respectively) the targets for these constraints should also be very similar. Choosing targets for these case-specific SQUEEZE constraints such that the average durations for the [1σ -v] and [2σ +v] cases turn out to be 132 ms and 131 ms is, however, difficult. **Crucially, they should not be set to 132 ms and 131 ms.** This is because these vowels are also subject to STRETCH[V] and SQUEEZE[V], which pulls them in a direction towards 198 ms, so the targets for the constraints specific to these constraints must be shorter, so that the observed durations are a compromise between the specific targets and the general targets.[26] As it happens, the targets for SQUEEZE[V/__σ] and SQUEEZE[V/__[-voice]] that predict distributions with averages in the right places are 52 ms and 73 ms, respectively.

In terms of harmony, the base case of [1σ +v] will incur violations from only the global constraints STRETCH[V] and SQUEEZE[V], while [2σ +v] and [1σ -v] will incur violations from these global constraints as well as from an additional SQUEEZE constraint, and [2σ -v] will incur violations from all four constraints. The violation functions for the constraints and the probability distributions that are predicted for all four of the inputs are plotted in Figure 25. The predicted means for each of these probabilities distributions is given in Table 12.

[26] This foreshadows the difficulty that will arise in discovering the ideal targets for constraints in these sorts of grammars: there is often no set of sample duration data whose average can be taken to be the target for a particular constraint, because the tokens in those data were subject to additional constraints as well, such that their observed durations represent a compromise between the constraint in question, and other constraints, whose targets also need to be found! This makes target learning a kind of hidden structure learning problem. As will be shown in Chapter 6, targets and weights can however be learned simultaneously in a way which maximizes the entropy of the data, providing a solution to this problem, though one which sometimes produces surprising results.

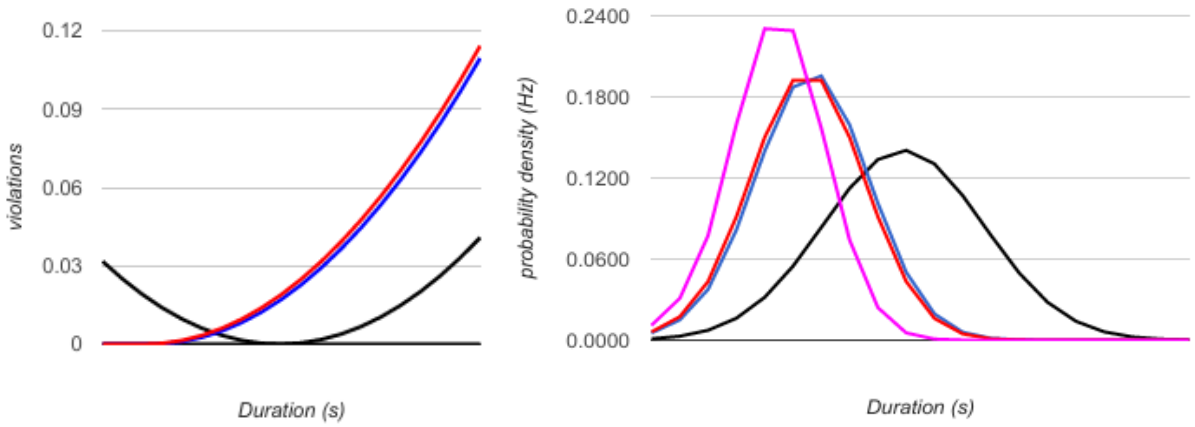

Figure 25: maxent duration à la Klatt. Left: the violation functions for STRETCH[V] + SQUEEZE[V] (black), STRETCH[V/__[-voice]] (blue), and SQUEEZE[V/__σ] (red). right: the probability distributions for [1σ +v] (black), [1σ -v] (blue), [2σ +v] (red), and [2σ -v] (magenta).

	1σ	2σ
/__[-voice]	(133 ms)	109 ms
/__[+voice]	(198 ms)	(130 ms)

Table 12: Predicted mean vowel duration, using the maxent grammar.

Since we selected the model parameters so as to match the mean vowel durations for the three simplest cases (as did Klatt), it is not surprising that the means for these cases roughly match those observed. However, given that no information was used about the [2σ -v] case where both effects apply, the model makes a striking prediction: the mean vowel duration in this case is predicted to be 109 ms, a much longer duration than predicted by models where each effect is treated as a multiplier, like Klatt's original model. Encouragingly, it is not far from the observed mean duration of 104 ms.

4.4.1. Constraint synergy

In fact, the duration of any input to this sort of grammar that undergoes multiple shortening processes will not be as short as predicted by a log-linear model fit to the results of observing each shortening process alone. This is in fact an emergent property of the maxent framework itself, and is true for the following reason: parabolic or hemi-parabolic shortening constraints, like SQUEEZE[V/__σ], apply the strongest "pull" on vowels which are the longest, i.e. the furthest from the target of the constraint. A segment in a grammar with few such SQUEEZE constraints, or in which these constraints have a low weight, will naturally be rather long. If a SQUEEZE constraint is added with a target much smaller than this length, the duration of such a segment will shorten in a dramatic way, reflecting the egregiousness with which its rather long default duration would have violated this additional constraint. However, as more SQUEEZE constraints with small targets are added, the amount of incremental shortening seen will each time be less dramatic, since it is already relatively closer to these targets, and the slope of a violation functions in the region close to its target is relatively shallower. In the extreme case, a shortening constraint, even a very highly weighted one, would end up having no discernable effect on the mean duration of a segment at all if that segment were already caused by the other constraints of the grammar to have a mean duration exactly at the target for the incoming constraint. In other words, constraint satisfaction in maxent phonetic grammars is predicted to be *synergistic* in the sense that a candidate can satisfy two constraints at once by moving towards both of their targets, moving less than would be expected if the effects of the two constraints were thought of as independent and straightforwardly additive.

Lengthening, of course, is similarly predicted to be synergistic. If two processes are assumed to be lengthening processes, the amount of lengthening that occurs when both apply is predicted

by maxent grammars to be less than would be expected if the effects were additive, since lengthening to fulfill one such constraint simultaneously helps to satisfy the other. For this reason, it is in fact crucial that the processes in the above example to be viewed as shortening processes, which apply to what in the unmarked case would be a more leisurely vowel, rather than lengthening ones, if the correct prediction about the fourth cell is to be made by a grammar fit to the other three. Treating the processes as monosyllabic lengthening and pre-voiced lengthening, modeled with STRETCH constraints on the relatively longer categories (Table 13), results in a grammar which massively under-predicts the duration of the [1σ +v] category (Figure 26), making predictions much worse than those of Klatt's baseline multiplicative model.

Pre-voiceless and disyllabic shortening		
Constraint	**Target**	**Weight**
STRETCH[V]	0.198 s	155
SQUEEZE[V]	0.198 s	155
SQUEEZE[V/__σ]	0.052 s	155
SQUEEZE[V/__[-voice]]	0.073 s	155

Pre-voiced and monosyllabic lengthening		
Constraint	**Target**	**Weight**
STRETCH[V]	0.109 s	155
SQUEEZE[V]	0.109 s	155
STRETCH[V/__(C)#]	0.142 s	155
STRETCH[V/__[+voice]]	0.135 s	155

Table 13: Grammars with two shortening constraints (left) and two lengthening constraints (right) based on the experimental results from Klatt, 1973b.

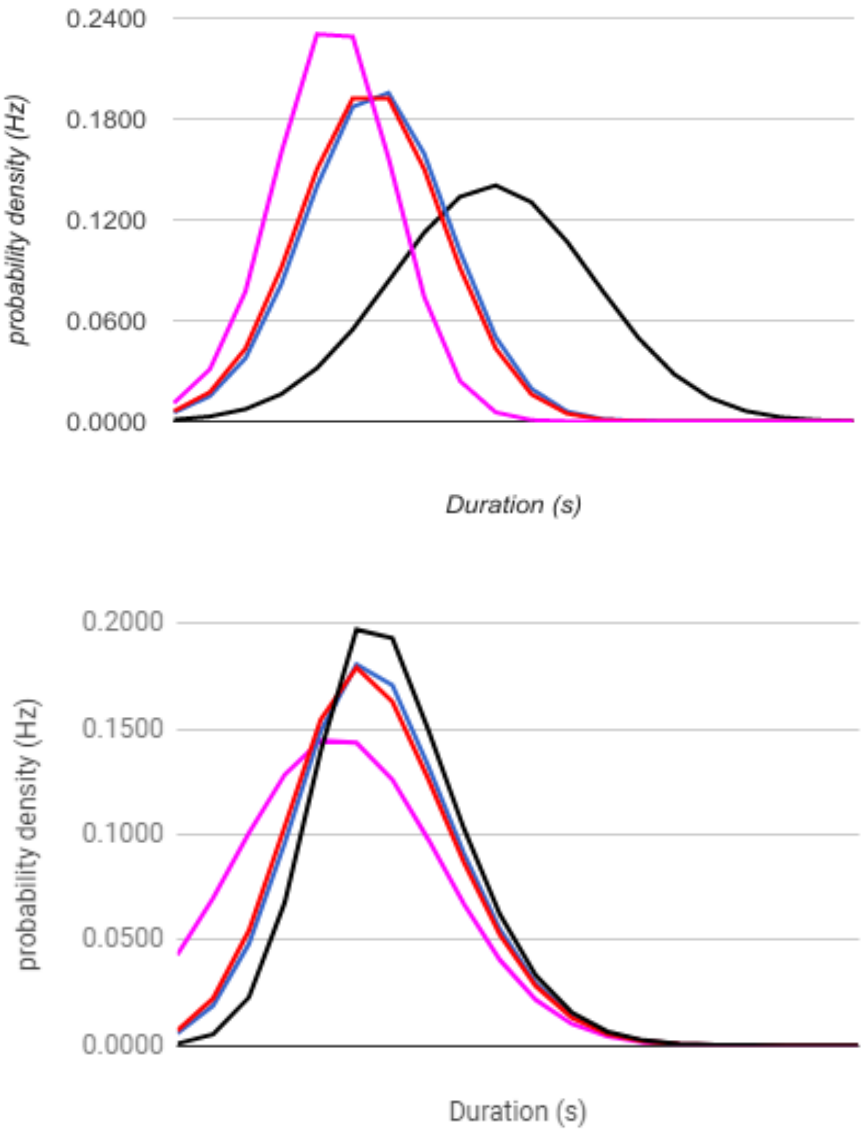

Figure 26: Probability distributions for the durations of vowels in the contexts [1σ +v] (black), [1σ -v] (blue), [2σ +v] (red), and [2σ -v] (magenta) in two maxent grammars, one where the two category-specific constraints are both SQUEEZE constraints and the longest category is the base case (top), and one where they are both STRETCH constraints and the shortest case is the base case (bottom).

4.4.2. Asymmetrical constraints and kurtosis

More careful investigation of Figure 26 reveals a set of empirical patterns regarding not only the means, but the shapes of the distributions of the four inputs to the grammar fragment described

in the previous section. The first is related to the discussion earlier in this chapter of the Consistent Variation Hypothesis: when more constraints apply to an input, its distribution will be narrower, meaning that less "random" variation is expected to be seen between tokens. In the grammar where the two effects are shortening effects, and the longest category (pre-voiced, monosyllabic) is treated as the "default" or "unmarked" case, this case is predicted to be the most variable, precisely because it is the least constrained, while the shortest category (pre-voiceless, disyllabic), being the most constrained, is predicted to be the least variable. Conversely, in the grammar where the two effects are lengthening effects, and the shortest category (pre-voiceless, disyllabic) is taken to be the default, it is predicted to be the most variable, and the longest category to be the least variable. This allows us, within the maxent framework, to make headway in characterizing individual effects on duration as either shortening or lengthening effects (or perhaps both), a distinction that, in many other frameworks, would be a completely arbitrary one.

The second, less obvious prediction is one that results from the use of the asymmetrical STRETCH and SQUEEZE constraints. In the shortening grammar, where category-specific effects are achieved with SQUEEZE constraints, the predicted distributions for the shortest three cases are not normal, but instead are slightly *negatively skewed*, lying in a distribution which is asymmetrical, having a longer tail with a shallower slope on the left side than on the right. The opposite is true for the longest three cases in lengthening grammar, which are *positively skewed*. This is because not all parts of all the distributions are governed by the same set of constraints: in the shortening grammar, for example, the sections of the distributions in the duration range shorter than 0.052 s are governed only by STRETCH[V] and SQUEEZE[V], since the SQUEEZE constraints exert no influence on candidates in the duration range below their targets. In the duration range longer than the targets of applicable SQUEEZE constraints, the slope of the

distribution will be steeper, because the combined weight of the constraints applying is greater, giving the overall distribution a negative skew.

This second result exposes another striking prediction of the maxent formalism: namely, that whenever two categories of segment vary only by whether or not they are subject to some SQUEEZE constraint (for example, [1 +V] as compared to [2 +V], with respect to the constraint SQUEEZE[V/__σ]), not only will segments subject to the constraint have shorter mean duration, their distribution should also have lower skewness, i.e. a steeper right side. STRETCH constraints are similarly able to result in higher skewness for the categories to which they apply. The result of all of this is that relatively longer categories—those to which more STRETCH and fewer SQUEEZE constraints apply, should have relatively higher skewness, if the grammar does indeed contain these kinds of asymmetrical constraints, and if these apply to only part of the range of realized durations of the candidate, as in the above example.

Note that the targets must lie somewhere within what ends up being the realized distribution in order for this effect on kurtosis to be observable. Consider, for example, some category of sounds is subject to a symmetrical DURATION constraint[27] (which, taken alone, produces a grammar that outputs a normal distribution), and additionally to an asymmetrical SQUEEZE constraint. When the target for SQUEEZE is similar to the target for DURATION, the distribution that results is no longer normal, but negatively skewed. This is because the left side of the distribution is more heavily constrained than the right side (Figure 27, top right). However, when the target for SQUEEZE is very small (set to 1, for the purposes of example), the predicted

[27] Or, equivalently, to a paired STRETCH and SQUEEZE constraint.

distribution is almost symmetrical, since the part of the distribution with non-negligible probability all lies above zero, and all of the parts of that duration range are subject to all the same constraints.

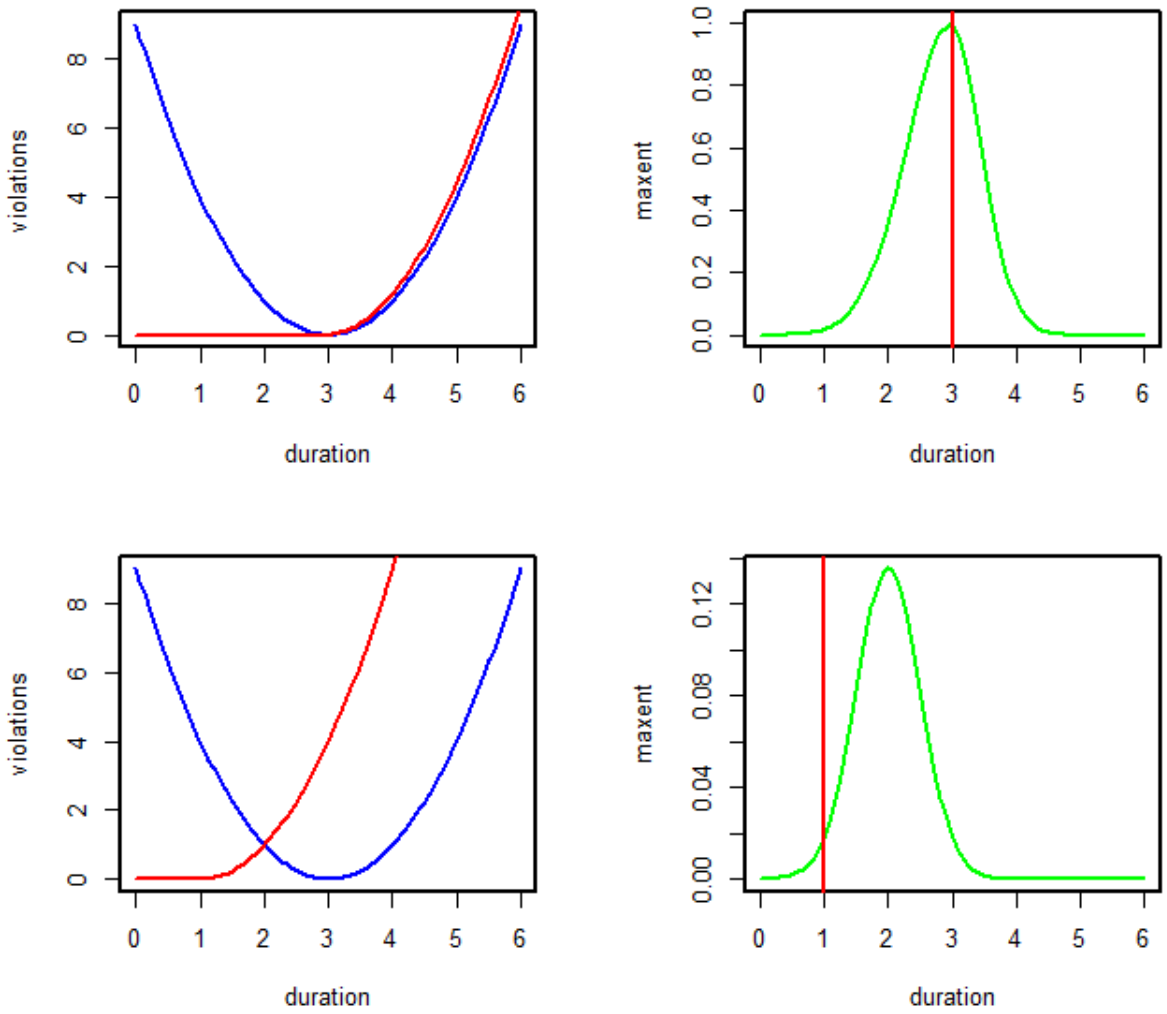

Figure 27: Violation profiles for DURATION (blue) and SQUEEZE (red), and predicted duration distributions (green) for grammars with the constraints DURATION and SQUEEZE. All weights are set to 1, the target for DURATION is set to 3, and the target for SQUEEZE is either 3 (top) or 1 (bottom), also shown as a verticle line in red.

In short: so long as the target values of asymmetrical constraints are not particularly extreme, those constraints are predicted to create asymmetries in the resulting distribution, such that additional

weight from SQUEEZE constraints leads to more negative skew, and vice versa for STRETCH constraints.

4.4.2.1. The Skewness Hypothesis

STRETCH and SQUEEZE constraints result in this effect on skewness while simultaneously contributing to differences in *mean* duration. As a result, languages generated by grammars with this constraint family should in principle show a correlation between mean and kurtosis: longer categories are necessarily subject to more STRETCH and fewer SQUEEZE constraints, creating, if anything, a more positive skew, and vice versa. This prediction is empirically testable, and is formalized here.

The Skewness Hypothesis

When one category of sounds (or larger prosodic constituents), is longer on average than some other comparable category, its distribution should have greater or equal skew to that other category, due to influence from more weight from STRETCH constraints, less weight from SQUEEZE constraints, or both.

In Chapter 5, this hypothesis will be tested experimentally, and moderate support for it will be found.

4.5. Summary of findings

Maximum entropy phonetic harmonic grammars are a promising tool for modeling the distributions of phonetic variables like duration, rather than just their means, and in fact make a number of predictions regarding how these distributions should behave.

One such prediction is that the amount of variation that a phonetic variable for a some category of sounds is observed to exhibit across phonological environments (i.e. its malleability in the face of external pressure) should match the amount of variation seen across tokens in a sample for this variable even within one phonological environments (i.e. its variability in response to random noise), because the same durational constraints constrain these two types of variation. This prediction, deemed the Consistent Variation Hypothesis, is empirically testable, and will be supported in Chapter 5.

Another prediction is that constraints which agree on a desired outcome will behave *synergistically,* in that candidates subject to both constraints will deviate less than would be expected if the effects of the two constraints were to be applied in an additive way. This property will be invoked in the discussion section of Chapter 5 to provide a potential explanation for the Hyperadditive Lengthening Generalization (Chapter 2, Chapter 5).

Two families of durational constraints, STRETCH and SQUEEZE, are explored, similar to the phonetic constraints proposed by Flemming (2001) in all respects except that they assign violations hemiparabolically instead of parabolically, asymmetrically penalizing only the candidates on one side of the constraint's target. This flexibility could be useful for implementing within maxent existing proposals in the phonetics literature which involve phonetic windows, the notion of maximum compressibility, and so on, without abandoning the many advantages of a harmony function which is locally continuous with respect to the phonetic variable in question. The STRETCH / SQUEEZE framework also results in the possibility of non-normality in the distribution of phonetic variables subject to multiple constraints, and further predicts that, in general, categories that undergo relatively more lengthening should also show higher skewness in

their distribution, while categories that undergo more shortening should show lower skewness: this fact is termed the Skewness Hypothesis.

Moving from HG to maxent also presents an implementational challenge, in that the normalization constant *Z*, which is required for the grammar to calculate probability distributions over candidates (a step necessary for maxent learning), is generally very difficult to calculate when the space of candidates is treated as continuous. However, an arbitrarily good approximation of the probability distribution can be made by discretizing the data: "binning" the candidates into durational ranges, and predicting a histogram rather than the probability density function proper. This approach will be taken in Chapter 6, where the parameters of example maxent grammars for duration will be fit to the experimental data from Chapter 5.

5. Data

The experiment described in this chapter investigates the effects on segment duration of a number of the independent variables described in Chapter 2, as well as what kinds of distributions appear for different segments in different segmental and prosodic contexts. The independent variables manipulated in these experiments all have main effects on duration that are already well documented, but the ways in which these variables actually interact to determine duration is somewhat less clear. Qualitative facts about the presence or absence of such interactions, as well as their directions, is of both empirical and theoretical interest.

While investigating duration using naturalistic corpus data, rather than experimentation, seems like an attractive option, it is problematic for several reasons: firstly, the corpus data would need to be annotated prosodically in a way that few corpora are, since intonation and prosody are so important for duration. Even if a natural corpus were annotated in this way, the coverage of the possible combinations of contexts would very likely be spotty—as Van Santen et al. (1997) put it, "due to the combinatorial complexity of any unrestricted language domain, even the most sophisticated text selection algorithms produce training text with disappointing coverage of the domain. Sparsity of the training corpus remains a central problem in duration data analysis." Rare combinations of conditioning phonological factors will almost certainly be missing or underrepresented. The experiment in this chapter has a Latin-square like design meant to avoid this sparsity, allowing the observation of many-way "interactions" between effects on duration. While not as naturalistic as other potential speech corpora, the highly targeted, fully intersective design of the experiments ensures that every possible phonological combination is represented in the training data, giving the learner better coverage even with a comparatively smaller training set, albeit one comprised of laboratory speech.

The results of the experiments are largely consistent with prior reports in the literature of interactions between various factors affecting duration (Chapter 2) and with the Hyperadditive Lengthening Generalization, the finding that most of these interactions are in the positive duration, such that the longest categories are longer than predicted by the main effects of these models, or, alternatively that the shortest categories are not as short as expected. In the discussion section, a number of potential explanations for this generalization are presented.

Enough data points are collected for each condition that an investigation can also be made into the shapes of the duration histograms for each experimental condition. This allows empirical testing of two empirical predictions made in Chapter 0: the Consistent Variation Hypothesis (section 4.1.5.1), a strong general prediction of the maxent phonetics framework, and the Skewness Hypothesis (section 4.4.2.1), a prediction of the asymmetrical STRETCH and SQUEEZE constraint family. Statistical tests suggest that both of these predictions hold for the experimental results.

In addition to answering empirical and theoretical questions about duration, the data collected double as training data for the phonetic learning algorithm developed in Chapter 6, serving as the input to learning for various grammar fragments for English front vowel duration.

5.1. Overview of the experiment

A production experiment was conducted in which five English front vowel phonemes were elicited in variety of segmental and prosodic contexts. The dependent variable was vowel duration (in statistical tests, log-duration was also modeled). The independent variables manipulated were the segmental features of the vowel, the number of consonants in the following coda (if any), which consisted of 0, 1, or 2 obstruents, the voicing of the following consonant (for the singleton codas

only), phrasal position, and pitch-accentedness, all of which are known to correlate with vowel duration.

In order to manipulate phrasal position and pitch-accentedness, the two prosodic factors, target words were elicited in four different sentential frames designed to encourage subjects to either focus or to background the target word, and to produce it either phrase-medially or phrase-finally.

While speech rate was not manipulated, it was treated as a random variable, and was estimated for each utterance on the basis of the duration of the carrier phrase in which the vowel was elicited.

5.2. Participants

49 UCLA undergraduates, 16 male and 33 female, were recruited from a research subject pool maintained by the UCLA Psychology Department consisting of students taking introductory psychology and linguistics courses. All were self-reported native speakers of English, and all lived in California at the time of the experiment. No dialect information was collected, but it seemed to the researcher and research assistants that nearly all of these students spoke some variety of California English. One participant's data did not sound like that of a native English speaker to the researchers—however, this speaker's data ended up being excluded on other grounds.

5.3. Methods and materials

5.3.1. Stimuli

32 monosyllabic target words were monosyllables selected so as to place five target vowels in a variety of syllable types and segmental contexts.

The five target vowels were the front vowels of English: /i/, /ɪ/, /eɪ/, /ɛ/, and /æ/. These vowels are known to show variation in their "inherent" phonetic duration. Phonologically, they vary in

height, in whether they are “tense” or “lax” (or peripheral or non-peripheral), and have varying degrees of diphthongization.

These five vowels were intersected with four codas: /d/, /t/, /ts/, and ∅: from these, the effects of voicing can be observed with /d/ and /t/, and the effects of a syllable’s being open, closed, and doubly closed on the duration of the vowel can be measured.

In order to produce more targets, the onset consonant was either /b/ or /m/, although this was not hypothesized to have a significant effect on vowel length.

The target words are given in Table 14.

	__Ø	__d	__t	__ts			__Ø	__d	__t	__ts
[i]	be	bead	beat	beats			me	mead	meat	meats
[ɪ]	---	bid	bit	bits			---	mid	mitt	mitts
[e]	bay	bade	bait	baits			may	made	mate	mates
[ɛ]	---	bed	bet	bets			---	---	met	---
[æ]	---	bad	bat	bats			---	mad	mat	mats

Table 14: Target words

Most of the “gaps” in space the onset / nucleus / coda combinations are the fault of English phonotactics: lax vowels do not occur word-finally in English. However, /mɛd/ and /mɛts/ are lexical gaps—while there are English words with these pronunciations, like ‘med’ (sometimes short for ‘median,’ ‘medical,’ ‘medieval’, etc.) and ‘Mets’ (the baseball team), these have lower frequency than the retained words, most of which were very common, and thus may not be familiar to all experimental subjects.

In order to manipulate the position of the target word within a frame, the target words were presented in two kinds of carrier sentence, one in which the target word occurs sentence-finally, and one in which it occurs medially.[28]

(27) Carrier 1: "No, [name] spelled [target word] correctly."

Carrier 2: "No, [name] correctly spelled [target word]."

The expectation was that, with the exception of the initial word 'no,' the entire sentence would be produced by most participants as a single intermediate phrase. As long as they did so, the target word would be both IP-medial and ip-medial in carrier 1, but IP-final in carrier 2. Furthermore, the two carriers are semantically and pragmatically very similar, and contain the same words and the same number of segments, to reduce the risk of confounds related to speech style, and to ensure that the length of the entire carrier sentence (excluding 'no') could be used as a proxy for speech rate.

Pursuant to this second goal, the proper names used in these sentences were all trochaic two-syllable names chosen from an online list of the most common English male and female given names. Since this name was always the first word in the part of the carrier phrase being used to estimate speech rate, only names beginning with an obstruent were included, in order to facilitate precise measurement of the timing of the left edge of this part of the carrier phrase.

[28] [name] stands for a semi-randomly chosen proper name—this part of the carrier was used to manipulate focus, as part of a method described on the following page.

Barry	Corey	Gladys	Shannon
Bertha	Curtis	Gordon	Sherry
Bradley	Debbie	Hazel	Stanley
Calvin	Derek	Hector	Tony
Cindy	Derrick	Herbert	Tracy
Clara	Dustin	Herman	Travis
Clifford	Florence	Jacob	Vernon
Connie	Francis	Jamie	Vincent

Table 15: Proper names used in carrier sentences.

In order to elicit both pitch-accented and non-pitch-accented tokens of each target word in each frame, the carrier sentences were presented as being the answer component of stimuli that consisted of question-answer pairs. The question component was a yes-no question of the form "Did [name] spell [filler word] correctly?" or "Did [alternate name] spell [target word] correctly?"

In the former case, the target word in the answer is new information, and all other information is discourse-given, encouraging the subject to focus the target word. In the latter case, the target word is discourse given and the *name* is discourse new, encouraging the subject to focus the name. Because focused elements in English are likely to receive nuclear pitch accent, and because the nuclear pitch accent cannot be followed by any further accents within the same intermediate phrase in English (Beckman & Pierrehumbert, 1986), the target word in this case would be unlikely to be unaccented.

To further encourage subjects to focus and accent the discourse-new word, this word was presented in ALL CAPS in the written stimuli.

With two carrier sentences and two question-answer contexts, a total of four frames were possible for each target word. These are summarized in Table 16.

	Unaccented	Accented
Final	Q: Did Bob correctly spell bed? A. No, SUSAN correctly spelled bed.	Q: Did Susan correctly spell fish? A: No, Susan correctly spelled BED.
Medial	Q: Did Bob spell bed correctly? A: No, SUSAN spelled bed correctly.	Q: Did Susan spell fish correctly? A: No, Susan spelled BED correctly.

Table 16: The four prosodic frames, using "bed" as the target word, and "Susan" as the proper name.

Since each of 32 target words could appear in four prosodic frames, a total of 128 phonologically distinct stimuli were created.

A short Python script was used to generate 128 question-answer pairs, randomly pairing names with target words and then inserting these pairs into the four question-answer frames above.

Only 64 of these stimuli were presented to each subject, in order to make the length of the experiment manageable. This was done by splitting the 128 pairs into two sets of 64. Each set contained exactly two stimuli containing each target word, so all participants encountered every target word. However, for any given target word, a given set either used that word in just the two accented frames, or just the two unaccented frames. Thus, no participant was ever asked to produce the same word (or proper name, for that matter) as accented in one utterance and unaccented in another utterance. This was done on the off chance that it would help to avoid confusion, as reading sentences in a prosodically natural way is already a difficult task.

5.3.2. Distractor items

In addition to the 128 stimuli, 96 question-answer distractor items were created. They were designed to elicit a variety of prosodic patterns, to keep subjects from defaulting to any one tune

over the course of the experiment. In keeping with the ‘spelling bee’ theme of the stimuli, most of the distractor sentences were related to that topic. Examples are shown in Table 17.

Q:	Did Audrey arrive on time?	A:	Yes, but Ida arrived late.
Q:	Does the teacher like Clyde?	A:	No, but she likes Dean.
Q:	How do you spell "dogs"?	A:	It's spelled D-O-G-S.
Q:	What did Jorge have to spell?	A:	Jorge had to spell "goal".
Q:	What letter comes after E in the alphabet?	A:	F comes after E.
Q:	What's a fruit that rhymes with "tango"?	A:	Mango.
Q:	What's a word starting with S?	A:	Well, "snakes" starts with S.
Q:	What's the etymology of "radius"?	A:	It comes from Latin.

Table 17: Examples of distractor items.

Each subject was given either the stimuli from Set A or from Set B interspersed with all of the distractor items. The order of the stimuli and distractors was randomized for each subject, with the constraint that all stimuli had to be separated by at least one distractor item.

Additionally, 10 warm-up items similar to the filler items were prepared, and prepended to the randomized stimuli and fillers for each subject. This was done in order to give subjects time to acclimate to the task and especially to establish a consistent speech style and speech rate before any data was collected.

5.3.3. Equipment

Subjects were recorded in one of two UCLA Phonetics Lab sound proof booths. Recordings were made using a head-worn Shure SM10A microphone, and run through an XAudioBox pre-amplifier and A-D device. The recording was done through PCQuirer, with a sampling rate of 11,000 Hz. Short sound files containing productions of individual stimuli were saved directly to disk in .wav format by a Matlab script.

5.3.4. Procedure

Stimuli were presented on the screen using a Matlab script. For each item, the question part of the question-answer pair was displayed in the center of the screen. Participants were asked to read this question silently. After an interval of 1.5 seconds, the answer appeared below it in red. At this point, text instructing the subject to read the answer aloud appeared in smaller print at the bottom of the screen. When the speaker pressed a key to continue, the screen was cleared and the next item was similarly displayed.

The Matlab script displaying the stimuli created individual recordings of the intervals between when the answer part of a stimulus was displayed and when the next stimulus appeared. These files were saved to disk with an appropriate filename generated by the script.

The first items were always the 10 warm-up items, after which subjects saw the 64 stimuli from either Set A or Set B interspersed with the distractor items in a semi-randomized order. At the exact midpoint of the experiment, the script displayed a screen advising participants to take a short break, allowing them to continue recording when ready by pressing the spacebar.

5.3.5. Forced alignment

Because the (intended) content of each utterance was already known, segmental alignment was done by using FAVE-align (Rosenfelder et. al. 2011). The alignments were then hand-checked for accuracy, although only the alignment of the segments of the target word and the location of the start and end of the intermediate phrase containing it (all of the utterance except ‘no’) were checked, since only the duration of these elements was being modeled. None of the automatic alignments needed to be adjusted on these grounds—the only times the aligner made significant

mistakes involved items where the subject had mispronounced or repeated parts of the stimuli, which were independently excluded from the data because of that disfluency.

5.3.6. Annotation and exclusion

3,136 individual productions were recorded. In order to be included, these utterances had to meet several criteria.

1. After the word 'no,' the utterance had to be a single prosodic phrase, with no perceivable juncture higher than word-level juncture. This was important to ensure phrase-mediality of the target words intended to be phrase-medial, and to make sure that speech rate, for which ip duration was used as a proxy, was measured consistently.
2. The target word had to have the accentuation / lack of accentuation expected by the researchers based on the frame it was in: the focus context had to be accented, and the background context had to be unaccented. The exact type of pitch accent produced was not controlled.
3. The utterance couldn't contain any disfluencies that would noticeably affect the overall duration of the phrase,[29] and couldn't contain any disfluencies at all in the target word or syllables adjacent to it.
4. If more than half of a speaker's utterances failed on the criteria already listed, the rest of their data was discarded as well.[30]

[29] Stuttered or repeated words were grounds for exclusion, for example, but mispronunciations of proper names were not, as long as the mispronunciation was also an obstruent-initial trochee.

[30] The reasoning here was that, since they probably didn't understand the task (or at least were not able to competently perform it), the phonetic naturalness of data from these subjects was generally not reliable.

Both the author and an undergraduate research assistant listened to all of the data and flagged utterances that failed to meet any one of these criteria; these utterances were then discarded. During this process, as soon as half of a subject's data were flagged for exclusion, the rest of their data was immediately excluded without being reviewed, to save time.

299 (9.5%) items were excluded on the first criterion, having a phrase-level juncture somewhere within the carrier sentence. 74 (2.4%) of these were due to the subject producing a phrase-level juncture between every two words in the stimuli.

460 (14.7%) were excluded on the second criterion, accentuation. Of these, 65 (2.1%) involved failing to accent the target word when it was supposed to be focused, and 395 (12.6%) involved accenting the target word when it was supposed to be backgrounded.

172 (5.5%) were excluded on the third criterion, for containing non-trivial disfluencies.

49 (1.6%) were excluded for failing on multiple criteria, and 28 (0.9%) were excluded for other reasons, such as clipping of the audio recording.

909 (29.0%) were excluded on the fourth criterion—19 subjects' data ware excluded wholesale in this way. Perhaps unsurprisingly, most of the unusable items were came from these same subjects: for many, reading aloud in a natural way proved quite difficult, especially when it came to prosodic focus. Many subjects chose to accent all words across the board, put phrase-level juncture between every pair of adjacent words, or both, treating the stimuli more like lists of words rather than sentences. Some failed to produce intonation patterns that were even remotely English-like, and instead seemed to have embarked on random walks through the pitch / duration / intensity space. Others had relatively more natural intonation, but had trouble reading aloud without frequent false starts or segmental disfluencies.

Ultimately, 1,917 utterances (61%) were excluded, and 1,219 (39%) were retained. While somewhat disappointing, the upshot to this is that the retained data are a fairly high-quality corpus: the remaining subjects understood the task and were able to produce natural-sounding speech in a laboratory environment, the retained data have reliable annotation for the prosodic properties that the experiment was designed to manipulate, and there are at least a handful of tokens for each possible combination of the independent variables.

5.3.7. Statistical tests

5.3.7.1. Linear Models

Two mixed effects linear regressions were run on the data, one with a dependent variable of vowel duration (the "linear" model), and one with a dependent variable of log duration (the "log linear" model). The log-linear model is in keeping with the idea that the factors influencing are multipliers, scaling the duration of a segment rather than addition to it, in line with the models proposed by Klatt (1973b) and Van Santen (1997), and others.

The independent variables were the ones manipulated in the study: the target vowel, preceding onset, following coda (if any), prosodic position, accentedness, and the length of the overall intonation phrase, which served as a proxy for speech rate (in the log-linear regressions, log IP duration was used). In both cases the fixed effects were binary feature representations of the onset, nucleus, coda, and prosodic condition, summarized in Table 18.

Onset Features	Nucleus Features	Coda Features	Prosodic Features
nasalonset (m_)	tense (i, eɪ) high (i, ɪ) low (æ)	closed (_d, _t, _ts) voiceless (_t, _ts) complex (_ts)	accented final

Table 18: Binary features used as fixed effects in the linear and log-linear regressions.

The duration of the whole intermediate phrase, which was the carrier phrase excluding the word “no” and excluding the target vowel, was used as a proxy for the speech rate of each utterance. This duration, termed “ip”, was included as a fixed effect in the linear model. The log of this duration, “ln ip”, was included as a fixed effect in the log linear model. A random effect of Subject was included in both models.

5.3.7.2. Interaction Effects

An observation was made in Chapter 2 regarding a tendency for effects on duration to interact positively with each other: the Hyperadditive Lengthening Generalization. In order to attempt to replicate the observations that lead to this generalization, a number of linear mixed effect regressions which included a single two- or three-way interaction were run, one for each interaction reported in the literature that involved factors that were part of the design of the current experiment.

5.3.7.3. Testing the Skewness Hypothesis

In Chapter 4, a prediction was made regarding the relative kurtosis of the probability distributions generated by the maxent formalism using the STRETCH and SQUEEZE family of constraints. The distribution of tokens subject to more SQUEEZE constraints should have both shorter means and smaller skewness, while the distributions of tokens subject to more STRETCH constraints should have both longer means and higher skewness.

The design of the stimuli (32 words and 4 prosodic contexts) were such that there were 128 experimental conditions (such as “meat” in accented final position, “beats” in unaccented medial position, and so on), and the durations for tokens from each of these subsets can be thought of as

a sample from a distribution. The mean and the skewness for each of the 128 samples was computed, and a Pearson's test for correlation was performed between the two sets of values.

5.3.7.4. Testing the Consistent Variation Hypothesis

A strong empirical prediction of the maxent phonetics framework was that categories of sounds which are relatively more constrained (subject an additional constraints or to a constraints with higher weight) should show relatively less variation in two ways: less "random" or unconditioned variation, and less variation in response to phonological factors, compared to a similar category of sound that is relatively less constrained: the Consistent Variation Hypothesis (section 4.1.5.1).

In order to test this prediction, we compared pairs of subsets of the experimental conditions which differ along only one phonological dimension (for example, comparing tense vowel data to lax vowel data, or open syllable data to closed syllable data), which in the maxent framework is the equivalent of observing the effects of just the constraint or constraints related to a single phonological factor. Each of these comparisons involved two sets of experimental conditions, which differ in one phonological dimension, but were pairwise matched along all the others.[31] The subsets compared, one for each of the nine binary phonological factors, are given in Table 19.

[31] This pairwise matching sometimes necessitated discarding some conditions in order to maintain a controlled comparison: for example, when comparing tense to lax vowels, all of the conditions with tense vowels in open syllables were omitted.

Feature	Conditions compared
nasalonset	All the conditions with /b/ onsets except those with target words /bɛd/ or /bɛts/[32] were compared to all the conditions with /m/ onsets.
tense	All the closed-syllable conditions with /eɪ/ or /i/ except those with target words /meɪd/ or /meɪts/ were compared to all of the conditions with /ɛ/.
high	All the conditions with /eɪ/ or /ɛ/ were compared to all the conditions with /i/ or /ɪ/ except those with target words /mɪd/ or /mɪts/, since the corresponding mid vowel categories, involving words of the form /mɛd/ and /mɛts/, were not present in the stimuli.
low	All the conditions with /æ/ except for those with target words /mæd/ or /mæts/ were compared to all the conditions with /ɛ/ (conditions with /eɪ/ were excluded from this set of mid vowel data since there were no low tense vowels).
closed	All the open syllable conditions were compared to all of the conditions with /t/-codas and tense vowels (conditions with lax vowels were excluded from the closed syllable set because lax vowels do not occur in open syllables, and /d/ and /ts/ data were excluded to eliminate effects of voicing and coda complexity).
voiceless	All of the conditions with /d/ codas except those with target word /mɛt/ were compared to all of the conditions with /t/ codas.
complex	All of the conditions with /t/ codas except those with target word /mɛt/ were compared to all of the conditions with /ts/ codas.
accented	All of the accented conditions were compared to all of the unaccented conditions.
final	All of the phrase-final conditions were compared to all of the phrase-medial conditions.

Table 19: For each of the binary features in the experiment, the pair-wise matched sets of experimental conditions which were compared to test the consistent variation hypothesis.

Each of these pairs of subsets differ in their mean duration. According to the maxent framework, those differences are necessarily the result of being subject to different constraints; for example, phrase-final vowel productions might be subject to a constraint DURATION[final], or

[32] Because there were no matched conditions with nasal onsets, since */mɛd/ and */mɛts/ were lexical gaps. These gaps were responsible for many of the other omissions mentioned in this table.

phrase-medial productions might be subject to a constraint like DURATION[medial], or both. One of the two categories might, depending on the constraints and their weights, be more heavily constrained than the other. This would result in a difference between the longer and shorter category in terms the amount of unconditioned variation seen within each of the individual experimental conditions in this category, but it should also result in a difference in the amount of phonologically conditioned variation, which can be estimated by observing the degree of variation between the means of the different conditions within the category.

The following example illustrates the method used to test this prediction, using the voiced coda and voiceless coda comparison.

First, comparable sets of experimental conditions corresponding to "voiced coda" and "voiceless coda" were collected, as described above. For each of these experimental conditions, the mean and standard deviation was computed. The mean of all the standard deviations was taken to be a decent estimate for unconditioned variation, i.e. how much pre-voiceless or pre-voiced vowels tended to vary due to random noise, averaged over a variety of phonological conditions. The standard deviation of the means across conditions was taken to be a decent estimate for phonologically conditioned variation, or how much the other phonological effects not related to voicing were affecting the duration of pre-voiced and pre-voiceless vowels. The delta of the unconditioned variation between pre-voiced and pre-voiceless is in an indicator of which category shows more unconditioned variation, and how much more. The delta of the conditioned variation between the two is an indicator of which category is more responsive to the other phonological effects. The maxent formalism predicts that the signs of these two deltas should generally be the same, since they both correspond to which of the two categories is more constrained by the grammar.

Pre-voiceless Conditions			Pre-voiced Conditions		
Condition	Mean (s)	SD (s)	Condition	Mean (s)	SD (s)
/bæt/, final, unaccented	0.196	0.044	/bæd/, final, unaccented	0.249	0.056
/bæt/, final, accented	0.209	0.036	/bæd/, final, accented	0.313	0.055
/bæt/, medial, unaccented	0.156	0.040	/bæd/, medial, unaccented	0.187	0.043
/bæt/, medial, accented	0.179	0.031	/bæd/, medial, accented	0.232	0.033
/bɛt/, final, unaccented	0.162	0.035	/bɛd/, final, unaccented	0.173	0.042
/bɛt/, final, accented	0.157	0.031	/bɛd/, final, accented	0.235	0.052
/bɛt/, medial, unaccented	0.124	0.029	/bɛd/, medial, unaccented	0.123	0.017
/bɛt/, medial, accented	0.153	0.031	/bɛd/, medial, accented	0.176	0.047
/beɪt/, final, unaccented	0.175	0.028	/beɪd/, final, unaccented	0.215	0.039
/beɪt/, final, accented	0.204	0.031	/beɪd/, final, accented	0.291	0.040
/beɪt/, medial, unaccented	0.154	0.026	/beɪd/, medial, unaccented	0.183	0.063
/beɪt/, medial, accented	0.178	0.023	/beɪd/, medial, accented	0.214	0.038
/bɪt/, final, unaccented	0.149	0.020	/bɪd/, final, unaccented	0.160	0.048
/bɪt/, final, accented	0.137	0.037	/bɪd/, final, accented	0.193	0.040
/bɪt/, medial, unaccented	0.107	0.030	/bɪd/, medial, unaccented	0.106	0.040
/bɪt/, medial, accented	0.114	0.025	/bɪd/, medial, accented	0.140	0.032
/bit/, final, unaccented	0.161	0.024	/bid/, final, unaccented	0.201	0.019
/bit/, final, accented	0.163	0.031	/bid/, final, accented	0.253	0.041
/bit/, medial, unaccented	0.125	0.025	/bid/, medial, unaccented	0.144	0.031
/bit/, medial, accented	0.148	0.019	/bid/, medial, accented	0.183	0.038
/mæt/, final, unaccented	0.162	0.052	/mæd/, final, unaccented	0.216	0.057
/mæt/, final, accented	0.170	0.042	/mæd/, final, accented	0.244	0.037
/mæt/, medial, unaccented	0.135	0.020	/mæd/, medial, unaccented	0.148	0.029
/mæt/, medial, accented	0.150	0.023	/mæd/, medial, accented	0.208	0.054
/meɪt/, final, unaccented	0.156	0.041	/meɪd/, final, unaccented	0.198	0.060
/meɪt/, final, accented	0.151	0.045	/meɪd/, final, accented	0.232	0.055
/meɪt/, medial, unaccented	0.108	0.011	/meɪd/, medial, unaccented	0.126	0.032
/meɪt/, medial, accented	0.135	0.022	/meɪd/, medial, accented	0.190	0.062
/mɪt/, final, unaccented	0.113	0.033	/mɪd/, final, unaccented	0.127	0.037
/mɪt/, final, accented	0.090	0.020	/mɪd/, final, accented	0.162	0.027
/mɪt/, medial, unaccented	0.090	0.019	/mɪd/, medial, unaccented	0.083	0.021
/mɪt/, medial, accented	0.084	0.027	/mɪd/, medial, accented	0.108	0.038

/mit/, final, unaccented	0.158	0.034	/mid/, final, unaccented	0.184	0.033
/mit/, final, accented	0.146	0.028	/mid/, final, accented	0.237	0.045
/mit/, medial, unaccented	0.116	0.019	/mid/, medial, unaccented	0.157	0.029
/mit/, medial, accented	0.128	0.023	/mid/, medial, accented	0.184	0.046

Unconditioned Variation (Mean of SDs)	0.029	Unconditioned Variation (Mean of SDs)	0.041
Conditioned Variation (SD of means)	0.031	Conditioned Variation (SD of means)	0.052

Δ Unconditioned Variation	0.012
Δ Conditioned Variation	0.021

Table 20: Means and standard deviations of the durations in the corresponding pre-voiceless and pre-voiced experimental conditions. Pre-voiceless conditions with no corresponding pre-voiced condition, namely those involving the targets "met," were excluded from the comparison.

Here, the pre-voiced conditions showed more unconditioned variation, and also more conditioned variation, than the corresponding pre-voiceless conditions. In the maxent framework, this suggests that the duration of vowels in the pre-voiceless category is more heavily constrained.

The same method was used to compare corresponding pairs of categories across of each of the other eight binary features, using the matched sets of conditions described in Table 19. The comparison was made using both duration and log duration.

5.4. Results

The number of tokens, means, standard deviations, and kurtoses for each of the 128 distinct categories are given in the appendix. Aggregate results are presented here.

5.4.1. Category means

As expected, mean vowel length was different for different vowels. Since the lax vowels only occur in closed syllables, rather than comparing mean durations of vowels across all tokens, the mean durations and mean log durations of vowels in just the tokens with closed syllables were compared (Figure 28). In this context, the front vowels from longest to shortest were /æ/, /eɪ/, /i/, /ɛ/, /ɪ/. As determined by two-tailed Welch t-test, the mean durations for each adjacent pair of vowels were significantly different ($p < 0.05$) as were the mean log durations for each adjacent pair.

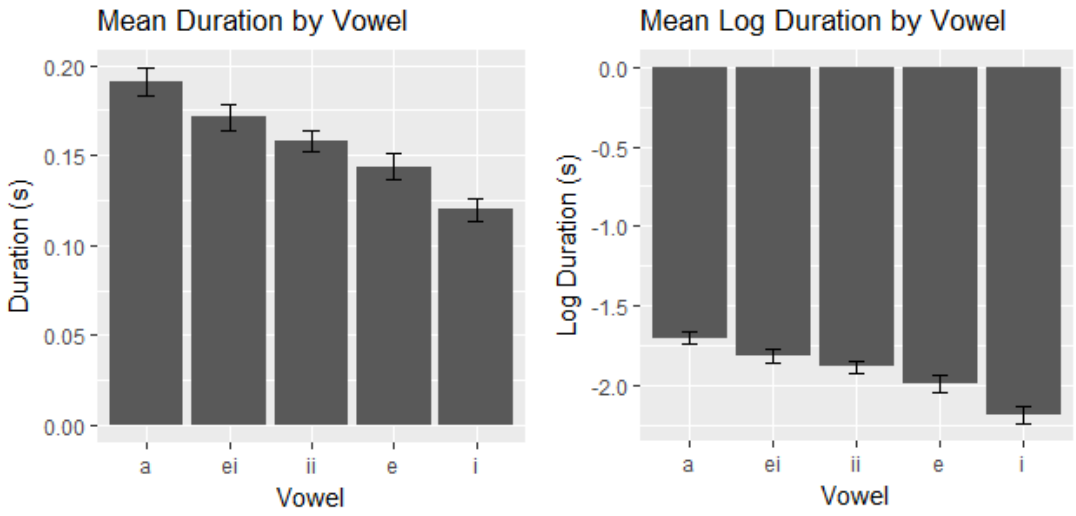

Figure 28: Mean duration and log duration by vowel phoneme for just the closed syllable data. Error bars represent 95% confidence intervals.

Mean vowel duration was different depending on the following coda, with open syllables being the longest, and closed syllables having length that varied by coda voicing and coda complexity. Since only tense vowels can be followed by all four coda types, /ts/, /t/, /d/, and Ø, means by coda type were compared for just the tokens with tense vowels, plotted in Figure 29 below. The codas associated with vowel duration from longest to shortest for this subset of the data were Ø, /d/, /t/, and /ts/. As determined by two-tailed Welch t-test, the difference in mean

duration and log duration of vowels before Ø and before /d/ was not significant ($p > 0.05$), while the mean durations and log durations between/d/ and /t/ and between /t/ and /ts/ were all significant ($p < 0.01$).

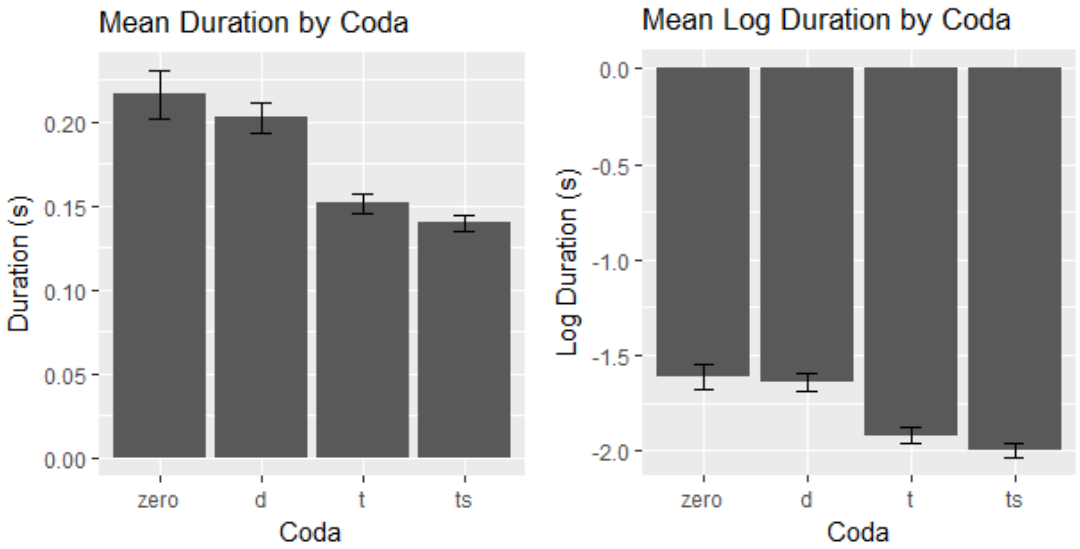

Figure 29: Mean duration and log duration by coda for just the tense vowel data. Error bars represent 95% confidence intervals.

To compare the effect of onset, since there were no stimuli of the form /mɛd/ and /mɛts/, the data with targets /bɛd/ and /bɛts/ were excluded. As determined by two-tailed Welch t-test, in this subset of the data the duration and log duration of vowels after /b/ was significantly longer than that of vowels after /m/ ($p < 0.001$).

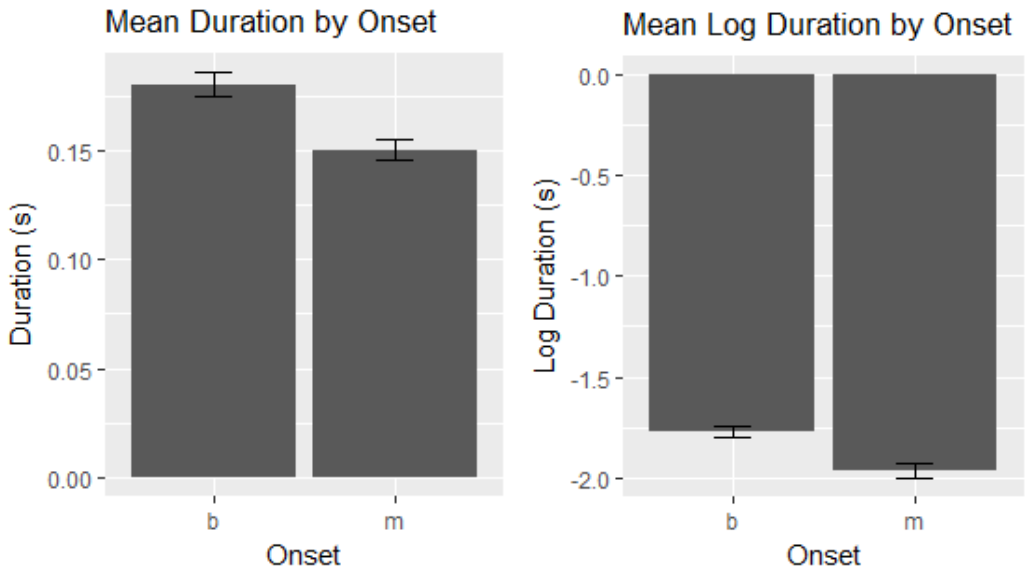

Figure 30: Mean duration and log duration by onset for all the data excluding targets "bed" and "bets." Error bars represent 95% confidence intervals.

The duration of vowels differed as a function of accentuation and phrasal position. As determined by two-tailed Welch t-test, the duration and log duration of vowels in the four prosodic positions were all pairwise significantly different ($p < 0.01$).

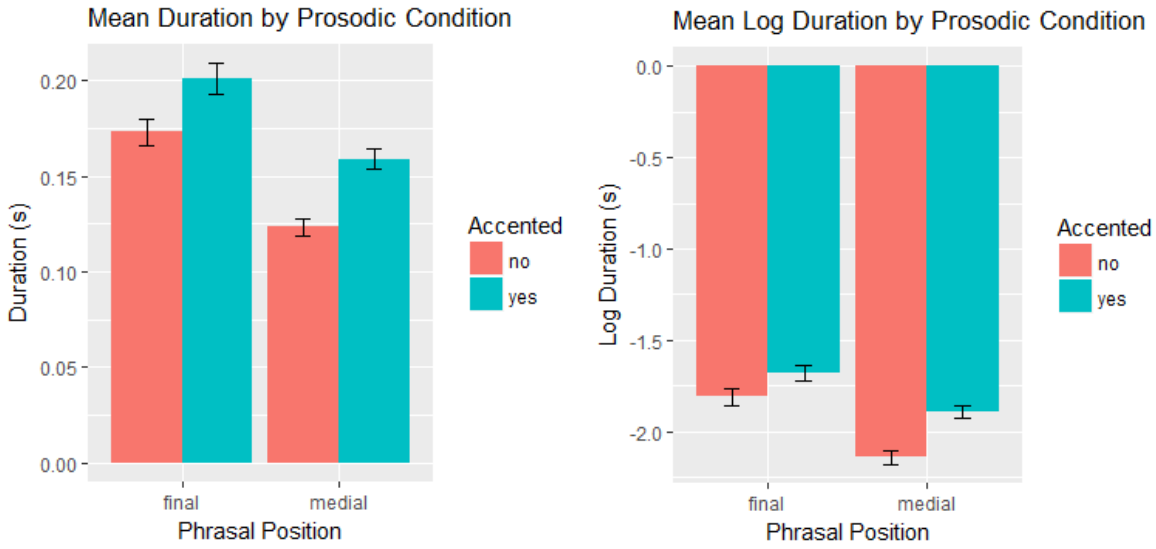

Figure 31: Mean duration and log duration by prosodic condition for all the data. Error bars represent 95% confidence intervals.

5.4.2. Linear and log-linear models

The effect sizes of the linear and log-linear main-effects only models are reported in Table 21. All of the fixed effects were significant ($|t| > 2$) for both models, and all were in the expected direction, with both degrees of vowel height inversely affecting duration, shortening effects of nasal onsets, closed syllables, voiceless codas, and complex codas, and lengthening effects of vowel tenseness, accentuation, phrase-finality, and longer ip duration (i.e. slower speech rate).

In terms of goodness of fit, the models performed similarly: the marginal R^2 values (Nakagawa & Schielzeth, 2013) for the linear and log-linear models were 0.59 and 0.61, respectively, with only slightly more variance being explained by the fixed effects of the log-linear model.

Fixed effects (linear model)

	Estimate	Std.Error	t-value
(Intercept)	0.098354	0.012043	8.167
nasal onset	-0.03621	0.002215	-16.346
tense	0.038589	0.002682	14.389
high	-0.014108	0.002424	-5.819
low	0.060058	0.003483	17.241
complex	-0.010537	0.00287	-3.671
voiceless	-0.04436	0.002792	-15.889
closed	-0.018801	0.003944	-4.767
accented	0.025913	0.002237	11.586
final	0.048054	0.002432	19.761
ip	0.043496	0.00714	6.092

Fixed effects (log-linear model)

	Estimate	Std.Error	t-value
(Intercept)	-2.08427	0.04272	-48.79
nasal onset	-0.23405	0.01314	-17.81
tense	0.28193	0.0159	17.74
high	-0.09383	0.01437	-6.53
low	0.39216	0.02067	18.98
complex	-0.0705	0.01703	-4.14
voiceless	-0.24572	0.01655	-14.85
closed	-0.05057	0.02339	-2.16
accented	0.15627	0.01326	11.78
final	0.26955	0.01457	18.5
ln ip	0.32383	0.06704	4.83

Table 21: The fitted parameters of linear and log-linear mixed effects regressions on the data, with a random variable of speaker.

5.4.3. Interaction effects

In Chapter 2, an interesting generalization was found in the literature with regards to the ways effects on duration interact: namely, a category undergoing any two lengthening effects (or avoiding any two shortening effects) was almost always longer than expected—this was termed the Hyperadditive Lengthening Generalization. Examples included positive interactions between phrase-finality and coda voicing (Umeda, 1975; Cooper and Danley, 1981), and vowel tenseness and coda voicing, as well as a three-way interaction between these two factors and vowel tenseness (Crystal and House, 1988). One negative interaction was seen, between phrase-finality and accentuation (Li & Post, 2014).

The data from the present study, in which many factors were independently varied, could potentially serve as a replication for some of these (and potentially many other) potential interactions between factors affecting duration.

For experiments with just a few fixed effects, it is sensible to include an interaction effect in the linear model and attempt to interpret the results. The present experiment, however, has 10 fixed effects: while in principle interactions between all of the effects could be tested for, this would create up to $2^{10} = 1024$ interactions (almost the number of data points), the model would by definition compute the means for each experimental condition perfectly, and the results would be meaningless. However, the two- and three-way interactions reported in the literature, and other similar simple interactions, can each be tested for independently, using regressions which contain only main effects and the particular interaction to be tested.

5.4.3.1. final × accented

Linear and log linear models identical to those above except that they included an interaction between *final* and *accented* showed a significant or marginally significant interaction (t = -1.25, t = -3.85). Accented vowels were less affected by phrase-final lengthening than were unaccented vowels, consistent with results from Li and Post (2014).

5.4.3.2. voiceless × accented

Linear and log linear mixed effects which included an interaction between *voiceless* and *accented* showed a significant interaction (t = -8.67, t = -6.92). As previously reported (De Jong, 2004; Choi et al., 2016), the effects of pre-voiced lengthening (or pre-voiceless shortening) were greater in accented than in unaccented words.

5.4.3.3. accented × low

Linear and log linear mixed effects which included an interaction between *low* and *accented* did not show a significant interaction between the two (t = 1.28, t = 0.80). However, the direction of the (non-significant) effects were both positive, consistent with the results from De Jong (2004).

5.4.3.4. voiceless × accented × low

Linear and log linear mixed effects which included an interaction between *low*, *accented*, and *voiceless* did not show a significant three-way interaction (t = 1.16, t = 1.57). However, the direction of the (non-significant) effects were positive, consistent with results from Choi et al. (2016).

5.4.3.5. final × voiceless

Linear and log linear mixed effects which included an interaction between *final* and *voiceless* showed a significant interaction (t = -10.19, t = -7.62). As previously reported (Umeda, 1975; Cooper and Danley, 1981; Crystal & House, 1988), the effects of pre-voiced lengthening (or pre-voiceless shortening) were greater in phrase-final than in phrase-medial position.

5.4.3.6. tense × voiceless

Linear and log linear mixed effects which included an interaction between *tense* and *voiceless* showed a significant or near-significant interaction (t = -2.54, t = -1.93). As previously reported (Crystal & House, 1988), the effects of pre-voiced lengthening (or pre-voiceless shortening) were greater for tense than for lax vowels.

5.4.3.7. Discussion

While not all the results reached significance in both models, the directions of all of the interaction effects were, encouragingly, consistent with those reported by prior authors, and therefore also all consistent with the generalization made in Chapter 2, with the exception of the negative interaction between phrase-finality and accentuation, the two phrase-level prosodic factors in the study.

If factors influencing duration tend to interact positively, we should also be able to see an indirect effect of this in the residuals of the main-effects-only models. In particular, many positive interactions should, all other things being equal, result in the data points predicted to be longest (i.e. those subject to multiple lengthening effects) being longer than predicted, or, equivalently, the data points predicted to be shortest (i.e. those subject to multiple shortening effects) being not as short as predicted.

As seen in Figure 32, exactly this pattern was observed for the residuals of the linear mixed effects model, but not for the log-linear mixed effects model, whose residuals were unpatterned with respect to predicted duration.

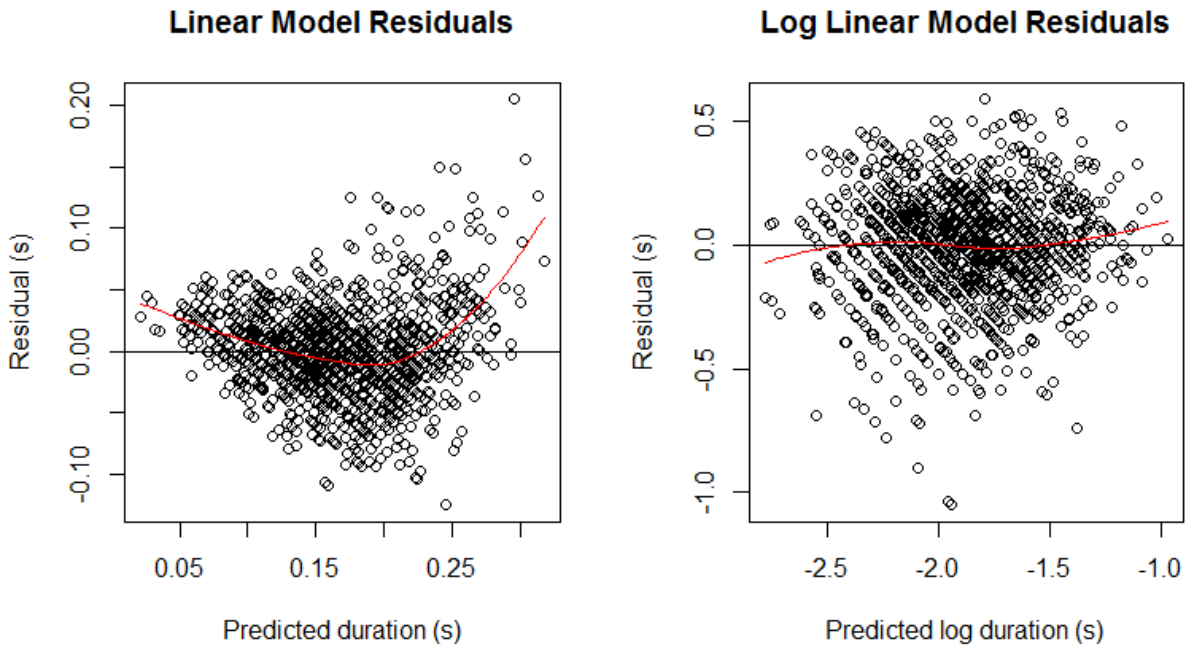

Figure 32: Residuals as a function of predicted duration for the linear (left) and log-linear (right) mixed effects models with only main effects, with loess lines.

This pattern suggests that part of the size of the interaction effects in the linear models can be explained by the fact that these models treat durational effects additively, when they should be treating them multiplicatively (or linearly in the log-domain), resulting in a number of apparent interactions which are really just artifacts of failing to transform the dependent variable appropriately. However, the fact that some interactions are present even in the log domain—i.e. the fact that effect sizes have been found to differ for different categories even when they are treated as proportions—is grounds to believe that this cannot be the full explanation.

5.4.4. Excursus on duration vs. log-duration as the dependent variable

Throughout this chapter, I have been ambivalent as to whether duration or log-duration should be the variable predicted. This uncertainty is related to a deeper question raised in Chapter 1 about how speakers perceive and represent the time dimension of speech: namely, whether they represent it in a way which is closer to linear, or closer to logarithmic (of course, it is possible that some other completely different transformation is applied instead).

From a quantitative perspective, the linear and log-linear models achieved about the same degree of model fit. However, some relevant qualitative observations can perhaps also be made. For example, as just discussed, the residual of the linear model shows a pattern which suggests that the vowel tokens predicted to be longest (those undergoing multiple lengthening effects) are longer than would be expected by a model which deals with duration in a non-logarithmic way, while the log-linear model shows no such patterned defect in its residual. The standard deviations across different vowels in different contexts are also more consistent when viewed in the log domain: if the data are split up into bins corresponding to the 128 experimental conditions (for example, "meat" in accented final position, "beats" in unaccented medial position, and so on), the 128 mean durations are significantly positively correlated with the 128 standard deviations ($r = 0.67$, $p < 0.00001$)—in other words, the contexts which produce longer vowels also have broader distributions. If an *a priori* assumption is made that different kinds of vowels in different sorts of phonological environments *should* show similar degrees of variation (and therefore similar standard distributions), the large positive correlation between mean duration and standard deviation could be taken as indirect evidence that speakers are not conceiving of the time dimension in a linear way, but in a proportional or logarithmic way, as argued by Rosen (2005). However, in the log-domain, mean log duration is in fact still significantly correlated with standard

deviation, just in the opposite direction: relatively longer categories show relatively less variance when viewed on the log scale ($r = -0.27$, $p < 0.005$), and this correlation could just as easily be seen as evidence that the data should not be transformed in this way.

Of course, the assumption that different phonological categories should show the same degree of phonetic variation is not a theoretically motivated one: differences in degrees of phonetic variation could well be linguistic in nature. Longer segments could be relatively more variable in the standard time domain for various reasons: speakers could for example have a bias towards more heavily constraining shorter categories. This pattern even makes a certain degree of practical sense: making even small changes to the duration of comparatively shorter sounds could have an adverse impact on the articulatability and perceptibility of the phonetic output, while changes of the same absolute size wouldn't effect long categories as adversely.

5.4.5. The shapes of durational distributions

While vowels with different segmental features or different segmental or prosodic environments showed differences in mean, they also showed differences in terms of the shapes of their observed distributions. As an example, Figure 33 shows histograms and probability density plots for all the tokens of /ɛ/ in unaccented phrase-medial position (relatively short vowels) and all the tokens of /eɪ/ in accented phrase-final position (relatively long vowels). In addition to having different mean durations (117 ms, 224 ms), the two subsets differ in standard deviation (29 ms, 80 ms), and in skewness (0.004, 0.648).

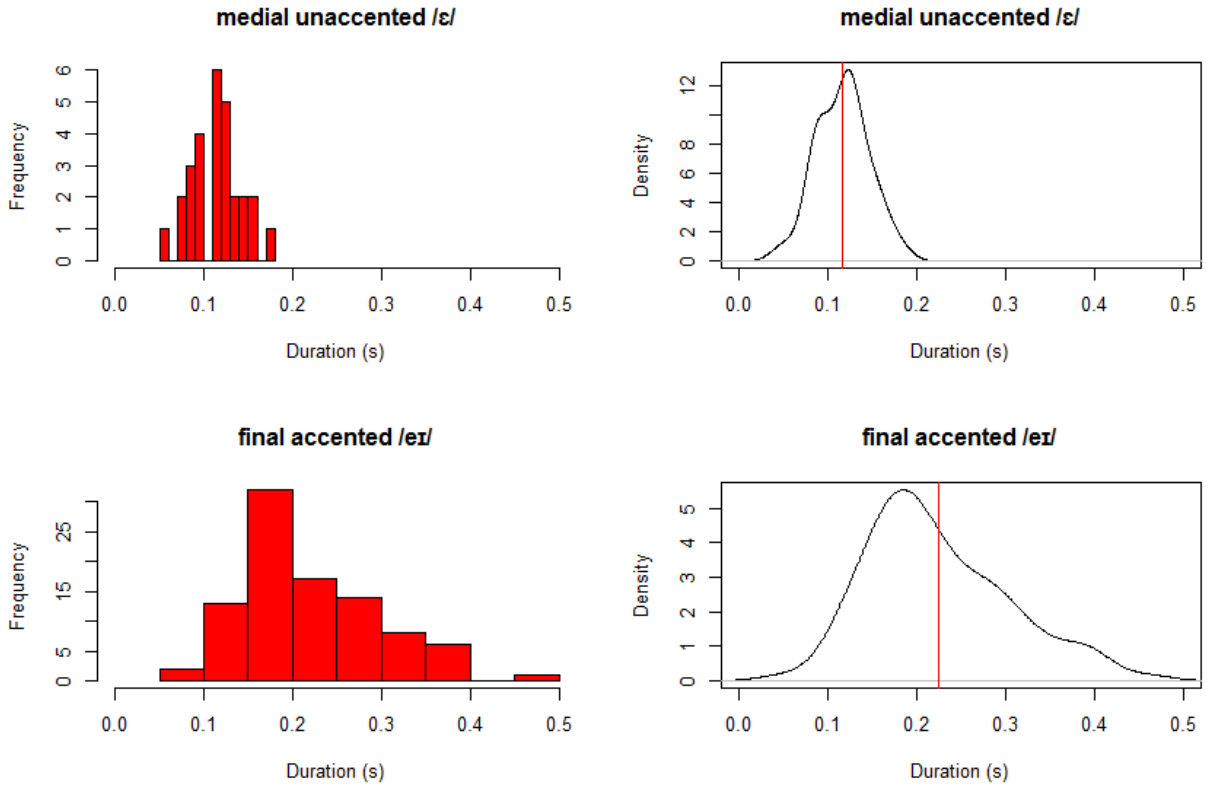

Figure 33: Histograms and probability density plots for tokens of phrase-medial unaccented /ɛ/ (top) and phrase-final accented /eɪ/ (bottom). Red lines indicate mean durations.

All this is of particular interest here because, in Chapter 4, symmetrical DURATION constraints were found to predict only normal distributions in maxent phonetic grammars, whereas constraints with asymmetrical violation profiles, such as STRETCH and SQUEEZE, were able to predict asymmetrical distributions. One specific prediction of these constraint families was that there could be a positive correlation between the duration of tokens in some phonological category and the skewness of the probability density function for that category. Another was a prediction of the maxent framework generally, independently of the constraint families used: a correlation between the unconditioned "random" variance in duration across tokens in some phonological category of segments, and amount of variation attributable to phonological effects. This was

because, in maxent, the very same constraints penalize random variance and phonologically conditioned variation, both of which involve deviation from constraint targets.

5.4.6. Testing the skewness hypothesis

The mean durations of the 128 samples were significantly positively correlated with their skewness ($r = 0.21$, $p = 0.016$). The mean log durations in each sample were similarly positively correlated with skewness of the log durations in that sample ($r = 0.20$, $p = 0.026$).

The results are plotted in Figure 34. These data are noisy, but this is to be expected: the number of tokens in each sample was relatively small (9.5 tokens, on average), such that the actual skewness of the sampled distribution can be estimated only very crudely. It is therefore striking that this correlation was found to be significant, even in the log domain.

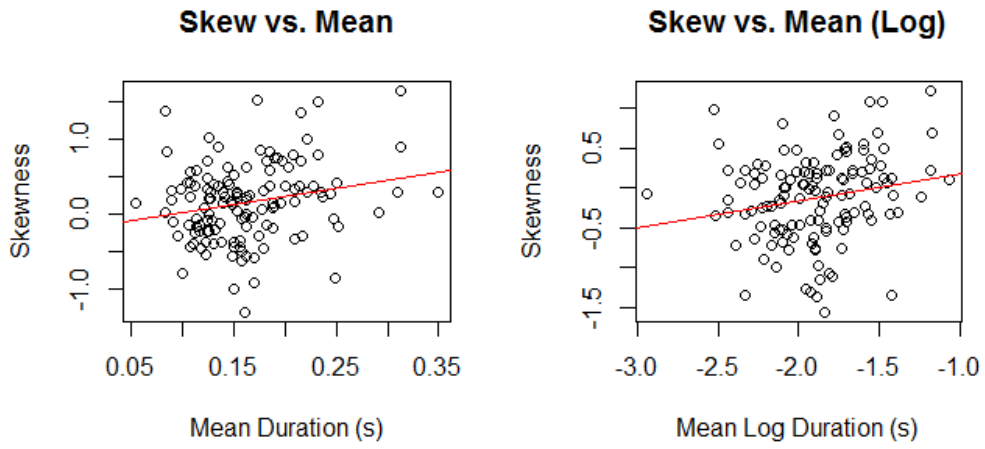

Figure 34: Skewness of samples as a function of their mean duration (left) or mean log duration (right) across the 128 experimental conditions.

5.4.7. Testing the uniform variation hypothesis

For the pairwise comparisons of subsets of the data, each made on the basis of one of the binary features used to model the experimental manipulations, the differences in the amount of

unconditioned variation between the longer and shorter subset of the data, as well as the difference in the amount of conditioned variation between the two subsets, is given in Table 22, and plotted in Figure 35.

Feature / Comparison	Duration		Log Duration	
	Δ Uncond. Var.	Δ Cond. Var.	Δ Uncond. Var.	Δ Cond. Var.
high (high vs. mid)	0.0027	0.0074	-0.0116	0.0043
low (ɛ vs. æ)	0.0063	0.0088	-0.0164	-0.0192
tense (ɛ, ɪ vs. eɪ, i)	0.0000	0.0036	-0.0559	-0.0601
nasal onset (m vs. b)	0.0006	0.0088	-0.0361	-0.0269
closed (t vs. Ø)	0.0146	0.0457	0.0083	0.1854
voiceless coda (t vs. d)	0.0116	0.0207	0.0068	0.0745
complex coda (ts vs. t)	0.0036	-0.0046	0.0165	-0.0500
accented	0.0051	0.0113	-0.0096	0.0180
phrase-final	0.0084	0.0198	0.0122	0.0320

Table 22: Differences between longer and shorter categories, across a number of phonological features, in their propensity toward conditioned variation, and toward unconditioned variation, for both duration (left) and log-duration (right).

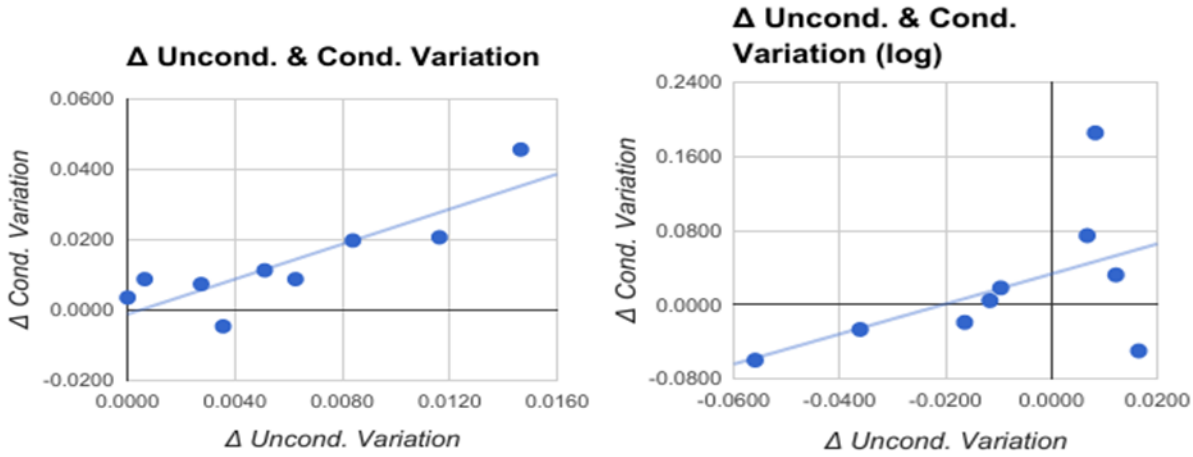

Figure 35: Differences between longer and shorter categories, across each of the nine binary phonological features, in their propensity to show conditioned and unconditioned variation, for both duration (left) and log-duration (right).

For actual durations, in all but one of the comparisons (coda complexity) the longer category (pre-voiced, phrase-final, etc.) showed both more unconditioned variation (in that the conditions in that category had larger standard deviations) and more conditioned variation (in that the means of the conditions had a larger standard deviation) than the corresponding shorter category (pre-voiceless, phrase-medial, etc). Furthermore, the degree to which the longer category showed more unconditioned variation was significantly positively correlated with the degree to which it showed more conditioned variation ($r = 0.86$, $p = 0.003$).

For the log durations, for six of the nine comparisons, the category which showed more conditioned variation also showed more unconditioned variation, but for the remaining three the direction of the difference differed. The degree to which the longer category showed more unconditioned variation and the degree to which it showed more conditioned variation were positively correlated, but the effect in the log domain did not reach significance ($r = 0.52$, $p = 0.162$). The exception is once again coda complexity (the data point in the lower right quadrant), an apparent outlier. If the coda-complexity comparison is discarded, the correlation becomes significant ($r = 0.76$, $p = 0.018$).[33]

This overall result is consistent with the Consistent Variation Hypothesis (section 4.1.5.1), an empirical prediction of the maxent phonetics framework.

[33] It could be relevant here that coda-complexity was entirely conflated with morphological complexity in this study: all of the branching codas were regular plurals, and these were the only polymorphemic target words.

5.5. Discussion and summary of findings

In line with prior literature, the duration of vowels were found to be affected by a wide range of factors, their segmental features, phrasal position, accentuation, speech rate, the manner of the preceding onset, syllable openness, and coda voicing.

More interestingly, vowels before the complex coda /ts/ were found to be significantly shorter than those before /t/—what Katz (2010) terms "incremental shortening." Incremental shortening, in which the vowel shortens in response to a non-adjacent consonant, is straightforwardly predicted by any model in which the durations of prosodic constituents like syllables are constrained, such that increasing the number of segments in a syllable should shorten the duration of each of them. Katz finds such incremental compensatory shortening effects for complex codas in general, but to different degrees, and did not find them at all for obstruent-obstruent codas like /st/, as compared to /s/ (/ts/ clusters were not tested). He explains some of the variation in the degree of compensatory shortening in terms of an asymmetry in the quality of perceptual cues for the vowel when the following consonant is a sonorant vs. when it is an obstruent, but notes that "even given this asymmetry, we might predict that obstruents induce less incremental CS for vowels, but we would still predict some. One possibility is that there really is a small effect, but the current study is not precise enough to uncover it; perhaps the effect is tiny in comparison to between-subject effects or random noise introduced by a failure to perfectly control for prosodic factors. In this case, there would be nothing left to explain." The present study seems to support this line of reasoning—the effect size found was very small (a difference of around 10 ms), and would not have been detectable had the prosodic factors, which had very large effect sizes, not been carefully controlled.

The linear mixed effects model (in which effects on duration were treated as additive) and log linear one (in which they were treated as multiplicative) achieved a relatively similar goodness of

fit. However, the linear model showed a pattern in its residual when plotted against predicted duration, whereas no such pattern was seen in the residual of the log-linear model, suggesting that the pattern seen in the former could be an artifact of failing to transform the data.

Consistent with a generalization found in the literature, there were significant interaction effects between the phonological factors, and almost all of these were positive interactions, in the sense that the longest categories were longer than expected, and the shortest not as short as expected. Possible explanations for this generalization are discussed in the immediately following section. The exception to this pattern was the interaction between the two prosodic factors, accentuation and phrase-finality, which showed a negative interaction—accented phrase-final vowels were not as long as predicted by the two main effects.

The skewness of distributions of duration across the 128 experimental conditions was correlated with their mean, suggesting that vowels subjected to multiple lengthening effects tend to be produced with distributions with higher skewness, and vice versa, consistent with the predictions of the maxent formalism combined with the STRETCH and SQUEEZE constraint family.

Finally, when comparing corresponding experimental conditions on the bases of a single phonological dimension, the subset of the data with more unconditioned variation tended to be the subset which showed more conditioned variation (variation induced by orthogonal phonological factors). Even though only 9 such comparisons were made, the correlation between conditioned variation and unconditioned variation was significant with respect to simple duration, and showed a similar trend for log-duration. That these two types of variation should in fact be correlated is a strong prediction of the maxent phonetics framework, regardless of the constraint set used, since more constrained categories of sounds should show less variation of all kinds, by definition.

5.5.1. The Hyperadditive Lengthening Generalization revisited

The results of the experiment in this chapter were consistent with the Hyperadditive Lengthening Generalization: a number of the interaction effects reported in the literature were confirmed to be significant, and for those that were not significant, the trend was in the right direction. The table from Chapter 2 summarizing these findings, to which the interaction results from the present study have been added, is reproduced below in Table 23.

	vowel features	coda features	coda complexity	lexical stress	accent	phrasal position
two-way interactions						
word-length (syllables)		Klatt (1975)				
vowel features		Crystal & House, 1988; Choi et al., 2016; (.)		De Jong, 2004	De Jong, 2004; Choi et al., 2016;	
coda features		Crystal & House, 1988	Katz (2010)	De Jong, 2004	De Jong, 2004; Choi et al., 2016; (*)	Umeda, 1975; Cooper & Danley, 1981; Crystal & House; (*)
coda complexity						
lexical stress					De Jong, 2004; Van Santen, 1992; Turk and White, 1999	Turk & Shattuck-Hufnagel, 2007
accent						Li & Post, 2014; (*)
three-way interactions						
lexical stress × accent	De Jong, 2004	De Jong, 2004				
vowel features × accent		Choi et al., 2016; (.)				

Table 23: Reported "interactions" between effects between factors affecting vowel duration. A (*) indicates that a significant interaction ($|t| > 2$) was found, and a (.) indicates that a trend ($|t| > 1$) was found, in the log-linear models of the experimental data from this chapter. All interactions were in the positive direction (if the effects are treated as both being lengthening or both being shortening effects), except for the interaction between pitch accent and phrasal position.

The observed tendency for durational effects to interact positively can probably be explained in a number of ways, but many will fall into two broad categories of explanation: "over-lengthening of the long" and "under-shortening of the short," depending on whether the longest data or the shortest data are taken to be the baseline.

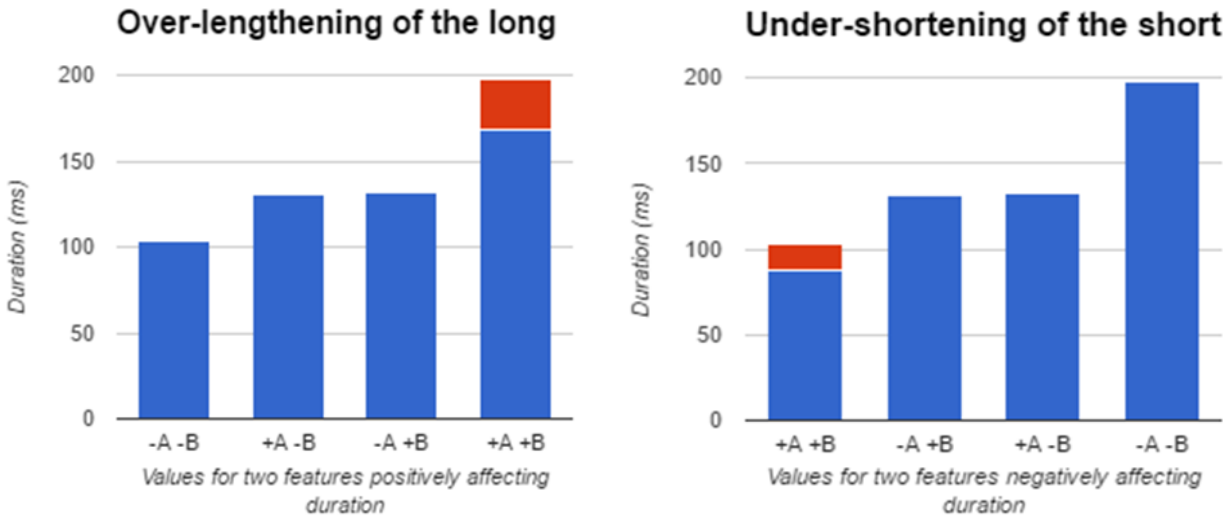

Figure 36: Two ways of understanding the Hyperadditive Lengthening Generalization. Blue bars indicate durations that would occur if the two effects combined multiplicatively with no interactions, where the longest duration (left) or the shortest duration (right) is predicted on the basis of the other three. Red bars indicate the deviation from this prediction that occurs for many pairs of durational factors, interpreted as unexpectedly long durations for either the category undergoing two lengthening effects (left) or two shortening effects (right).

5.5.1.1. Over-lengthening of the long[34]

One explanation for the generalization is that categories of sounds (and perhaps larger prosodic constituents) which are already long, due to some phonological feature or other, are affected more by additional lengthening processes than are comparatively shorter categories. In this case, the longest of the four categories, in which two lengthening factors are taken to apply together, comes out longer than expected, and is the "outlier" which causes there to be an interaction.

One example of a theoretical motivation for a pattern of this sort, at least with respect to the interactions that in some way involve prosody, is the relationship between prosodic prominence

[34] Or, equivalently, "under-lengthening of the short."

and listener-oriented hyperarticulation. De Jong (1995), for example, describes prosodic prominence as "localized hyperarticulation," positing that material that is prosodically more prominent (stressed or focused syllables and words), because it is the most important for successful communication, is produced in a way which maximizes its communicative potential. One component of this hyperarticulation is to maximally differentiate phonological contrasts, some of which are cued by phonetic duration. Other accounts (e.g. Aylett & Turk, 2004) also treat prosodic prominence, and the associated lengthening effects, as (at least in part) a perception-enhancing strategy.

To explain why this predicts hyperadditive lengthening of the long, consider the following hypothetical example. One of the cues distinguishing /æ/ from /ɛ/ is the fact that it is a good deal longer. When placed in a prosodically prominent position (a position of localized hyperarticulation), this contrast is best preserved by lengthening /æ/ quite a bit, but *not* lengthening /ɛ/ by as much, since doing the latter would depreciate the durational cue that distinguishes the two vowels.

For a phonetic constraint-based approach which explicitly inhibits lengthening of an inherently short category of segment based on output-output correspondence, see Braver (2013).

5.5.1.2. Under-shortening of the short[35]

A second explanation for the generalization is that categories which are already short, due to some phonological feature or other, are less affected or unaffected by additional shortening processes. In this case, the shortest of the four categories, in which two shortening factors are taken

[35] Or, equivalently, "over-shortening of the long."

to apply together, comes out longer (not as short) as expected, and is the "outlier" which causes there to be an interaction. At least two existing models of duration, as well as the phonetic constraint model laid out in Chapter 4, include a mechanism which directly or indirectly predict that already short segments will shorten to a lesser degree.

Klatt (1975), after finding interactions of this sort, introduced into his duration model a "minimum duration" parameter for each category of speech sound, based on its segmental identity. His multiplicative (log-linear) model was changed to compute from the remaining factors not absolute duration, but duration above this minimum duration baseline. Say, for example, that the default duration of the vowel /i/ was 100 ms, its minimum duration was 90 ms, and the effects of stress and phrase-final lengthening were both multipliers of 150%. When only one effect applies, the vowel would be 90 ms + 10 ms * 1.5 = 105 ms, but when both apply, it would be 90 ms + 10 ms * 1.5 * 1.5 = 112.5 ms. So the values in a 2x2 table describing the results of the two factors would be 100 ms, 105 ms, 105 ms, and 112.5 ms. The non-zero minimum duration baseline, then, effectively results in hyperadditive lengthening of the long: when applying final lengthening, unstressed /i/ goes from 100 to 105, an apparent 5% increase in duration, while stressed /i/ goes from 105 ms to 112.5 ms, an apparent 7.1% increase in duration.

Katz (2010), along the same lines, includes a mechanism for enforcing minimum vowel duration in his phonetic constraint grammars. While constraints governing the durations of segments and syllables are at first assumed to have parabolic violation functions, following Flemming (2001), to account for a lack of shortening in some cases, these violation functions are taken to behave parabolically only above a certain minimum duration threshold, but to assign infinite violations below that threshold, effectively enforcing a minimum duration for the segment or syllable being constrained. A result of this mechanism is that if the constraints governing two

particular duration effects are such that the shortest category of sounds would (in the simple parabola model) fall below some minimum duration threshold, the output for that shortest category will be longer than otherwise expected: essentially, there are floor effects on duration.

A different explanation for under-shortening of the short—one which does not involve minimum duration—was seen in Chapter 4. In that chapter, constraints were seen to behave *synergistically* in the phonetic maxent framework: when a segment had already shortened in response to one constraint, the effect of a second shortening constraint was weaker, because the segment was already relatively close to the second constraint's target, and the slope of the parabolic violation function in that range was shallow. An unshortened segment would react more to the second constraint, because its duration would be further from the second constraint's target, and the slope of the parabolic violation function in that range would be steep.

Compounding two lengthening constraints on relatively longer categories should, according to this theory, also behave *synergistically*: segments should lengthen less to each additional lengthening constraint introduced. This, however, is the opposite of the pattern of interactions seen in most of the data. The Hyperadditive Lengthening Generalization is therefore only explicable in this framework by way of synergistic shortening, and not synergistic lengthening. In other words, if all or most of the effects in question are shortening effects enforced by constraints on the shorter of the two categories differentiated by any given effect (with the possible exception of either accentuation or phrase-final lengthening, which negatively interact with each other). For example, the effect that has been described by some pre-voiced lengthening would need to be recast as pre-voiceless shortening, with vowels in pre-voiceless position being the ones subjected to an additional constraint. Likewise, relatively shorter vowel phonemes would need to be governed by more or more heavily weighted constraints, and longer vowel phonemes by fewer.

If lengthening and shortening processes do behave synergistically, it would therefore be necessary to assume that the "default" durations of segments and other prosodic constituents are in most cases reflected by their longest possible durations, and that the various segmental and prosodic factors which effect duration are all (or mostly all) compression effects, curtailing what would in the unmarked case be more leisurely productions.

This idea that shorter categories are generally more constrained, in addition to being an explanation for hyperadditive lengthening, is right in line with another empirical result from this chapter: that the longer of two subsets of sounds was also generally the one which showed more variation—both more random variation, and more variation in response to phonological conditioning.

In Chapter 6, in which a maxent learner is trained on the experimental data, it is found that, at least for grammars involving symmetrical DURATION constraints, the learner does in fact assign more weight to constraints on relatively shorter categories, and that constraint sets which include only these shortening constraints perform better than those which also include lengthening constraints.

6. Maxent phonetic learning

This chapter describes an algorithm for the maxent learning of the parameters of phonetic constraint grammars for duration, given a corpus of training data. It is an adaptation of the phonological maxent learning algorithm described by Goldwater and Johnson (2003), modified to accommodate the aspects of phonetic grammars which differ from their phonological counterparts.

Using this algorithm, the weights and/or targets of several different sets of constraints, including symmetrical DURATION constraints (Flemming, 2001) and STRETCH and SQUEEZE constraints (Chapter 4), are trained on the experimental results from Chapter 5, producing models (grammar fragments) for the phonetic duration of English front vowels in a variety of segmental and prosodic environments.

The learner is much more successful in creating grammars which fit the training data when it is allowed to simultaneously learn targets and weights. Additionally, certain constraint sets are found to be superior to others. For the best of these, the grammars produced by the learner seem to fit the data well, modeling both variation in the means of different categories of sounds in different contexts, and variation in the shapes of the samples' distributions, a capability special to the maxent variety of phonetic constraint grammars.

6.1. Learning in the phonetic domain[36]

For phonological maxent learners (Goldwater & Johnson, 2003; Hayes & Wilson, 2008), the goal of the grammar is to predict, as closely as possible, the observed frequencies of some set of

[36] Hayes and Schuh (MS) have concurrently adopted a very similar formalism for phonetic constraints on duration (in particular, the durations of syllables in the *rajaz* meter of Hausa), and likewise use maxent learning

surface representations, which correspond either to possible realizations of some underlying representation (alternation learning), or all potential surface forms in the language (phonotactic learning). The linguist must provide the learner with a finite set of constraints,[37] a finite[38] set of candidates for each input, and violation profiles of the candidates with respect to the constraints—in other words, OT tableaux. The model parameters to be learned are the weights of the constraints. In order to learn these weights, the learner makes use of an "objective function," which in maxent is a function that computes the log-likelihood of the training data given some setting for the constraint weights. The learner learns optimal weights, i.e. those which maximize the log-likelihood (or "entropy") of the training data, via gradient descent—in other words, it starts with arbitrary weights, and repeatedly takes steps in a direction which is "downhill" from the perspective of the objective function, until it finds an optimum. In order to determine which direction is downhill, the objective function computes not only the probability of the data, but also the "gradient" or slope at the present location with respect to the constraint weights.

The task of learning maxent grammars for phonetics, while somewhat analogous, is more complicated for several reasons. The first complication involves the nature of the phonetic candidate space. Phonetic variables are continuous, such that the set of possible outputs, rather

---

to fit the weights of these constraints (though not the targets, which are determined using aggregate statistics) to a corpus of Hausa song.

[37] Or, in the case of algorithms which tackle the problem of "constraint selection", a set of possible constraints, from which an algorithm selects a subset to include in the grammar (Hayes & Wilson, 2008). This kind of approach is in some sense mathematically equivalent to training a grammar on a very large constraint set, but with the requirement that most constraints should have weights of zero.

[38] While GEN, in theory, is usually assumed to generate an infinite number of candidates, linguists in practice do not make use of infinite tableaux, instead generally opting to exclude candidates which will demonstrably be assigned vanishingly small probabilities by the grammar. In phonotactic learning, this is sometimes accomplished by only considering word forms shorter than a certain number of segments.

than being a discrete set, is better thought of as a vector space. As discussed in Chapter 3 and 4, while grammars could in principle operate on this vector space directly, the normalization step necessary for maxent learning may not be practically feasible. A practical solution to this is to discretize or "bin" the phonetic candidate space into equally sized intervals (say, into candidates with durations up to 10 ms, 10-20 ms, 20-30 ms, and so forth), making the candidate set once again finite, as discussed in Chapter 4. This simplification makes the task at hand more analogous to phonological learning (at least from the perspective of the learning software). While this "rounding" of phonetic variables could have some effect on the grammar learned, this effect can be minimized by reducing the size of the intervals, such that the results of learning would match those of a continuous learner arbitrarily well.

The second, more significant complication is that phonetic constraints have targets in addition to weights,[39] doubling the number of model parameters. In phonological learning, each candidate incurs a predetermined number of violations of each constraint, and the constraints' violation profiles can be thought of as the features[40] of the training data—the problem of learning constraint weights then becomes analogous to the problem of learning the best coefficients for the features of a logistic regression, as pointed out by Goldwater and Johnson (2003). By comparison, if constraints also have target parameters, the constraints are not properly features in this way until the target parameter has been set—DURATION[V, 40ms] and DURATION[V, 45ms], for example, are effectively two completely different features in the sense that they produce different

[39] Though see Windmann et al., 2015, for a proposal involving phonetic constraints without targets.

[40] The term 'features' is here used in the machine learning sense of being dimensions of the *x* input vector to the model, not in the linguistic sense of phonological features.

violation profiles. Target learning can therefore be thought of as a special case of constraint selection, or model selection.

For the linguist, two approaches to this complication can be taken. The first approach is to adopt a strategy for selecting or estimating targets first, based on the properties of the learning data, prior to the maxent learning step, which is (as in phonological learning) only used to learn the weights of constraints. The second approach is to simultaneously learn the targets and the weights for some set of constraints, using maxent learning. In the latter approach, the maxent learning step is much less straightforward, because the learning space (the shape of the objective function to be minimized through gradient descent) becomes qualitatively quite different from the well-behaved, convex space that occurs when only weights need to be learned.

Versions of the learning algorithm which implement each of these learning approaches will be presented in this chapter, along with the results of applying them to the experimental data from Chapter 5 to produce grammar fragments for English front vowel duration.

6.2. The learning algorithm

This section describes an algorithm for learning model parameters for maxent phonetic constraint grammars. The user provides to this learner training data (like the data collected in the experiment in Chapter 5), constraint definitions, and, optionally, constraint target values. The learner finds the constraint weights (and the constraint target values, if not provided) that maximize the likelihood of the training data.

6.2.1. Constraint definitions

Phonetic constraints are formalized as functions from candidates (themselves I/O pairs, where the input is a set of phonological features, and the output is a duration) to violations. Constraints

have two hand-selected components: a criterion defining which sounds are to be constrained based on their phonological properties,[41] and a "polarity" of either 1 (STRETCH), -1 (SQUEEZE) or 0 (DURATION), the last case being a symmetrical constraint with parabolic violations, of the sort used by Flemming (2001) and others. The other two components of each constraint—its weight and its target—are the parameters to be learned or estimated. The unparameterized constraint SQUEEZE[+high], for example, would be defined as the tuple <nucleus ϵ {i, ɪ}, -1>, and the fully parameterized constraint SQUEEZE[+high, w=2, t=0.8 ds] as the four-tuple as <nucleus ϵ {i, ɪ}, -1, 2, 0.8>.

6.2.2. Constraints

All of the constraints employed in this section are constraints on vowel duration, and constrain either all vowels, or a natural class defined by exactly one of the nine features used to categorize the experimental data (Table 18): *high, low, tense, nasal onset, closed, voiceless coda, complex coda, accented,* and *phrase-final*. For each of these features, both the positive and negative values define a natural class, so a total of 1 + 9 + 9 = 19 natural classes can be constrained in this way. Two families of constraints on duration are used: the symmetrical DURATION constraints (Flemming, 2001), and corresponding pairs of asymmetrical STRETCH and SQUEEZE constraints (Chapter 0), for a total of 57 possible constraints (Table 24).

Treating all constraints as constraints on vowel duration is a drastic simplification, since some are "standing in" for what are probably in fact constraints on larger prosodic constituents like

[41] For the purposes of the current data, this criterion itself had two components: a feature or column in the training data, like "nucleus" or "phrasal position," and a set of values that an input needed to have in that column in order to be targeted, such as {i, ɪ}, getting the natural class [+high], or {final}, getting vowels in phrase-final syllables.

syllables or even words—closed syllables, for example, probably contain shorter vowels because of interplay between constraints on syllable duration and constraints on vowel duration (Katz, 2010), rather than a contextually limited constraint on vowel. Unfortunately, the more principled approach taken by Katz and others involves a high-dimensional set of candidates, in which all of the sounds in a syllable are simultaneously governed by the grammar, which the learner described in this chapter is not quite up to handling at the time of writing, and for which there is perhaps not adequate training data in any case, since the durations of all of the sounds in the syllable would need to be measured accurately. However, constraints like DURATION[V/closed] can be thought of as modeling what would, in a more principled grammar, be the violations contributed to DURATION[σ] by the duration of a vowel in a closed syllable, above and beyond what is already contributed by DURATION[V].

	DUR vowel	SQUEEZE vowel	STRETCH vowel
Constraints on longer classes	DUR non-high	SQUEEZE non-high	STRETCH non-high
	DUR low	SQUEEZE low	STRETCH low
	DUR tense	SQUEEZE tense	STRETCH tense
	DUR open	SQUEEZE open	STRETCH open
	DUR singleton coda	SQUEEZE singleton coda	STRETCH singleton coda
	DUR pre-voiced	SQUEEZE pre-voiced	STRETCH pre-voiced
	DUR post b	SQUEEZE post b	STRETCH post b
	DUR accented	SQUEEZE accented	STRETCH accented
	DUR phrase-final	SQUEEZE phrase-final	STRETCH phrase-final
Constraints on shorter classes	DUR high	SQUEEZE high	STRETCH high
	DUR non-low	SQUEEZE non-low	STRETCH non-low
	DUR lax	SQUEEZE lax	STRETCH lax
	DUR closed	SQUEEZE closed	STRETCH closed
	DUR complex coda	SQUEEZE complex coda	STRETCH complex coda
	DUR pre-voiceless	SQUEEZE pre-voiceless	STRETCH pre-voiceless
	DUR post m	SQUEEZE post m	STRETCH post m
	DUR unaccented	SQUEEZE unaccented	STRETCH unaccented
	DUR phrase-medial	SQUEEZE phrase-medial	STRETCH phrase-medial

Table 24: A superset of the constraints used in any one learning attempt.

The whole set of constraints wasn't used in any one training run: instead, different combinations of these constraints were used in an effort to determine which set of constraints was best able to capture the data (sections 6.4 and 6.5).

6.2.3. Tableaux construction

The algorithm first automatically generates a set of tableaux. The input for each tableau is a surface representation, representing some vowel in some phonological context. The candidate outputs are phonetic duration ranges or "bins," with widths equal to a duration resolution specified by the linguist, and with the shortest and longest bin corresponding to minimum and maximum

durations, also specified. Duration is represented in deciseconds (tenths of a second), rather than seconds or milliseconds, throughout the algorithm.[42]

Each tableau is first assigned a "basic" violation profile (Table 25)—a representation of which constraints are active at all in that tableau, given the phonological properties of the surface representation input. For example, while a candidate duration for the vowel in the word "cat" may incur violations of SQUEEZE[+low], depending on the duration of the candidate and of the constraint target, it will never violate SQUEEZE[+high].

The algorithm then reads in the training data, and sorts them into tableaux, one for each distinct surface representation. For each tableau it computes counts for each of the duration bin candidates, corresponding to how many data points with the relevant surface representation fall into each duration range. It also computes a basic violation profile for each tableaux (Table 25).

[42] This was done because it was found that using seconds resulted in very small violations, such that the weights learned were very inconveniently large, while using milliseconds had the opposite problem. For example, a candidate with duration 0.1 s (100 ms) violating a STRETCH constraint with a target at 0.3 s (300 ms) incurs 0.04 violations when represented in seconds, and 40,000 violations when represented in milliseconds. The optimal weights change correspondingly—they will be exactly 1,000,000 times smaller in the latter case. In both cases, the large discrepancy between the magnitudes of the weights and the magnitudes of the targets results in unnecessary amounts of information loss due to rounding, but also creates a situation where the optimal area in the weight/target search space used by the gradient descent function is very narrow in some dimensions, and very large in others, potentially making it more difficult for the gradient descent function to converge.

[mit] accented, phrase-final	counts	STRETCH [V]	SQUEEZE [V]	...	STRETCH [+low]	SQUEEZE [+high]
i = 0 ds	0	1	1	...	0	1
i = 2 ds	1					
i = 4 ds	3					
i = 6 ds	2					
i = 8 ds	1					
...						
i = 20 ds	0					

Table 25: Hypothetical example of a tableau prepared during initial tableaux construction, prior to learning, with duration range 0 - 500 ms and duration resolution 20 ms.

If targets are to be learned simultaneously with weights during the maxent learning step, this is the extent of the pre-processing of the data that can be done before learning, and the input to the learner is this set of basic tableaux. If targets are pre-selected by the linguist, these tableaux can be augmented with actual violation profiles for each candidate based on the constraint targets and the candidates' durations (for an example of what this would look like, see Table 26).

6.2.4. The objective function

As in phonological maxent learning, optimal parameters can be found by minimizing an objective function (or "cost function"), which is the negative log likelihood[43] of the training data (given some arbitrary parameter values), plus a prior on the constraint weights.

[43] While maximizing the likelihood and maximizing the log likelihood are equivalent, working in the log domain is computational faster, since sums are faster to compute than products, and less prone to rounding error, since many of the maxent values, being inverse exponentials, are vanishingly small.

The likelihood of the training data is simply the product of the predicted probability of all of the data points, and the log likelihood is therefore the sum of the logs of the predicted probabilities of all of the data points.

Starting from the initial tableaux, and from some set of parameter values for the constraint weights and targets, the learner first computes actual violations for each cell in each tableaux, using the hemi-parabolic violation functions described in Chapter 4. After this, the log-likelihood is computed the same as it would be for phonological maxent grammars: overall harmonies are computed for each candidate, the negations of these harmonies are exponentiated, and the resulting numbers are normalized so that they add to one, and can be interpreted as probabilities (Table 26). The logs of these probabilities are multiplied by the counts, and summed, giving the log probability of the tableau in question, and this process is repeated for all the tableaux to find the log likelihood of the entire corpus of training data.[44]

[44] It is common practice when using gradient descent to, whenever possible, compute the gradient of the objective function with respect to the model parameters, so that the gradient descent algorithm knows which direction is downhill, rather than having to estimate. While computing the gradients for the weights is unproblematic, and already part of the implementation of existing phonological maxent learners (e.g. Hayes & Wilson, 2008), computing the gradient for the targets proves much more difficult, in part because changing a target changes the probability of the observed data points both directly, in that it changes their harmony, and indirectly, in that it changes the harmonies of other candidates as well, which changes Z (the normalization factor), which in turn changes all the output probabilities. Therefore, a gradient descent function which estimates the gradient is used here.

[mit] accented, phrase-final	counts	STRETCH [V] w = 2 t = 5 ds	SQUEEZE [V] w = 1 t = 2 ds	...	STRETCH [+low] w = 1 t = 10 ds	SQUEEZE [+high] w = 2 t = 4 ds	Harm-ony ($\hat{w} \cdot \hat{v}$)	Pr (e^{-h})	Prob (Pr/Z)
i = 0 ds	0	5^2	0			0	25	< 0.000000001	< 0.00000001
i = 2 ds	1	3^2	0			0	9	0.00012	0.018
i = 4 ds	3	1^2	2^2			0	5	0.0067	0.98
i = 6 ds	2	0	4^2	...	0	2^2	20	0.000000002	0.0000003
i = 8 ds	1	0	6^2			4^2	52	< 0.000000001	< 0.00000001
...		...	...			...	...	...	
i = 20 ds	0	0	18^2			16^2	580	< 0.000000001	< 0.00000001
								Z = 0.0069	

Table 26: Hypothetical example of a maxent tableau created during learning.

The *prior* cost was defined to be the sum of the constraint weights multiplied by a small number, 0.0001.[45] The effect that the prior has on learning is to slightly penalize larger constraint weights, such that, all other things being equal, the learner prefer models with smaller weights. Constraints which are redundant or have no effect on candidate probabilities, which would otherwise have no optimal weight, will with the addition of the prior be assigned a weight of zero. If the same fit to the data could be accomplished either with one constraint with weight *w*, or with two constraints with combined weight > *w*, the prior will cause the learner to favor using a single

[45] This prior is linear with respect to the constraint weights, where many implementations of maxent learning use a Gaussian prior in which each constraint weight is squared before the constraint weights are summed. Using a Gaussian prior, a high penalty is incurred when any one constraint has a large weight, such that, when multiple constraints have the same or a similar effect, the learner will prefer to spread weight across all these constraints to minimize the sum of their squared weights, even when some of those constraints are completely redundant to the model. By contrast, the linear prior will simply pick the constraint that works the best, and assign a weight of 0 to constraints that are not needed. The linear prior was used here because it made it easier to see in the results which constraints were redundant.

constraint. While strong priors can often used to bias a learner towards simpler models at the cost of optimizing the likelihood of the data (for example to model linguistic learning biases, or to prevent overfitting), the coefficient chosen here for the prior is so small that maximizing likelihood is treated as the first priority, and model simplicity as a secondary one, such that the prior practically only has an effect when multiple sets of parameters predict the data equally well.

6.2.5. Parameter learning

Given the objective function just defined, optimal model parameters—just the weights if targets are provided by the linguist, or both the weights and the targets simultaneously—can be estimated using gradient descent.

Parameter estimation was done using the function scipy.optimize.fmin_l_bfgs_b from the Python library SciPy (version 0.19.0, downloadable from https://www.scipy.org as of 6/3/17).

The bounds for the constraint weights were 0 and 1000, precluding negative or extraneously large weights. When targets were to be learned as well, target values were bounded at -50 and 50 ds (-5.0 and 5.0 s).[46] Additionally, the targets of STRETCH constraints were given a lower bound of the minimum candidate duration, 0 ds, and the targets of SQUEEZE constraints were given an upper bound of the maximum candidate duration, 5.0 ds (500 ms). If the targets for these asymmetrical constraints stray outside of these bounds, the constraints would stop penalizing any candidates at all: this lack of any violations would then make the local gradient for the target value completely flat (since adjusting them slightly in one direction or the other would wouldn't affect the likelihood of the data) which could prevent the learner from converging.

[46] While the optimal constraint targets were initially expected to be positive, or at least 0, this assumption was not forced upon the learner, and in fact turned out to be incorrect.

Initial weights for the constraints were random numbers between 0 and 0.2,[47] and initial targets were set to the average duration of the sample data (1.66 ds), plus a random number between -0.2 and 0.2 ds.

The precision parameter of the fmin_l_bfgs_b function was set to 10.0 (very accurate).

6.3. Criteria for assessing models

To assess the relative merit of the models learned in the following section, I use Akaike's Information Criterion (AIC), a measure of model informativity based on the log-likelihood of the data according to the model (its goodness of fit), and the number of degrees of freedom that model has. I take the number of degrees of freedom of the maxent model to be the number of constraint weights, plus the number of constraint targets when these targets are parameters set by the learning algorithm, but not when they are pre-selected by the linguist in some deterministic way, based on aggregate model statistics.

6.4. Weight learning with pre-selected targets

As already discussed, if the linguist (or human learner) can find a way to provide appropriate targets first, the constraint weight learning step becomes the same as in phonological learning, since candidate violation profiles are available at the outset of the learning step. The problem, of course, is how to select good constraint targets.

[47] Using larger initial weights resulted, in the first few learning steps, in violent and oscillatory changes in the target values, which in turn had the effect of causing the learner to set the corresponding weights to 0, where they sometimes stayed for the remainder of the training run.

One simple and perhaps intuitive option is to use aggregate statistics over the training data: if vowels in open syllables are, on average, 216 ms in duration, then perhaps this is a good approximation of the target duration for the constraint or constraints on vowels in open syllables. From an acquisition perspective this strategy seems plausible as well: upon observing a number of tokens for some class of sounds or prosodic constituents, or tokens of sounds in some phonological context, the human learner would have access to their average duration, and, other things being equal, attempt to achieve that duration for phonologically similar tokens in their own grammar.

A constraint grammar consisting of all 19 DURATION constraints (set D_{full}) with target values set to be the mean duration of the tokens governed by the constraint in question, was trained on the data. Multiple training runs, with different random initial starting values for the weights, were conducted, and all of them converged to the same values. This is expected, because learning space of just the constraint weights is *convex*, in that it has no local optima, and there learner will therefore not get stuck anywhere that is worse than the global optimum (Della Pietra et al., 1997). The learning results are reported in Table 27.

The learned model is very sparse in its use of constraints. Though no weight is given to the universal DUR[vowel], all of the 128 input SRs are at least subject to one constraint,[48] avoiding any pathological "flat" predicted distributions. However, putting no weight on either DUR[low] or DUR[non-low], the model does not differentiate at all between /æ/ and /ɛ/, which is clearly a defect, since in reality these phonemes have very different mean durations (191 ms and 143 ms, respectively).

[48] All SRs with tense vowels are constrained by DUR[tense], and lax vowels are constrained by either DUR[singleton coda] or DUR[complex coda].

constraint	weight	target (ds)	constraint	weight	target (ds)
DUR high	0.05	1.50	DUR non-high		1.77
DUR non-low		1.60	DUR low		1.91
DUR lax		1.55	DUR tense	0.32	1.81
DUR closed		1.59	DUR open		2.16
DUR complex coda	0.33	1.39	DUR singleton coda	0.38	1.68
DUR pre-voiceless	2.00	1.42	DUR pre-voiced		1.93
DUR post m		1.51	DUR post b	0.67	1.79
DUR unaccented	0.26	1.48	DUR accented		1.79
DUR phrase-medial	0.86	1.44	DUR phrase-final		1.89
DUR vowel		1.66			
Neg. Log Prob.	**3695.40**				
AIC	**7428.79**				

Table 27: Learned weights for constraint set D_{full} with targets pre-set to the means of the natural class they constrain. Weights of 0 are omitted for visual clarity.

Perhaps, however, the failure of this grammar is to be expected. While initially intuitive, the mean duration of a natural class is, for these sorts of grammars, almost certainly not the best value for the target of a constraint on that natural class. Even in the very simple toy example given in Section 4.4, the targets constraints specific to vowels in disyllabic words and vowels before voiceless codas had to be set to vowels *shorter* than the means of these categories: this was because the actual realized duration was the result of not only these constraints, but of other constraints as well. In that example, the other constraints to which they were subject had targets *longer* than the means of these categories, so the category specific constraints had to have targets *shorter* than their means in order for the grammar to get the desired result.

To test whether this might be the case in the more complex grammar currently at hand, two additional training runs were performed with targets estimated in a different way. For the DURATION[vowel] constraint, the target was still set to be the mean, but for the constraints on

longer than average categories, like DUR[low], the target was set to the mean plus either one or two standard deviations, and for the constraints on shorter than average categories, like DUR[non-low], the target was set to the mean minus either one or two standard deviations.

The model using targets adjusted by 1 SD in this way (Table 28) achieve a quantitatively much better fit than the original grammar, and the model using targets adjusted by 2 SDs (Table 29) achieves an even better fit. In addition, both assign a non-zero weight to DURATION[low], distinguishing between mid and low vowels. The 2 SD model, however, does not have any constraint which distinguishes between vowels in syllables with [t] vs. [ts] codas, missing a pattern which is empirically present in the data, despite achieving an overall much better fit.

constraint	**weight**	**target (ds)**	**constraint**	**weight**	**target (ds)**
DUR high	0.19	0.91	DUR non-high		2.42
DUR non-low		0.96	DUR low	0.06	2.51
DUR lax		0.94	DUR tense	0.79	2.45
DUR closed		1.02	DUR open	0.66	3.01
DUR complex coda	0.45	0.98	DUR singleton coda	0.16	2.29
DUR pre-voiceless	1.56	1.00	DUR pre-voiced		2.60
DUR post m	0.18	0.92	DUR post b	0.09	2.43
DUR unaccented	0.52	0.94	DUR accented	0.55	2.46
DUR phrase-medial	0.92	0.96	DUR phrase-final	0.18	2.59
DUR vowel		1.66			
Neg. Log Prob.	**3454.81**				
AIC	**6947.62**				

Table 28: Learned weights for constraint set D_{full} with targets pre-set to the mean (black), the mean plus 1 SD (red), or the mean minus 1 SD (blue) of the natural class they constrain. Weights of 0 are omitted for visual clarity.

constraint	weight	target (ds)	constraint	weight	target (ds)
DUR high	0.25	0.31	DUR non-high		3.06
DUR non-low		0.33	DUR low	0.12	3.11
DUR lax		0.34	DUR tense	0.48	3.09
DUR closed	0.33	0.45	DUR open	0.82	3.86
DUR complex coda		0.57	DUR singleton coda		2.90
DUR pre-voiceless	1.43	0.57	DUR pre-voiced	0.33	3.27
DUR post m	0.21	0.32	DUR post b		3.08
DUR unaccented	0.47	0.40	DUR accented	0.61	3.13
DUR phrase-medial	0.94	0.47	DUR phrase-final	0.24	3.29
DUR vowel		1.66			
Neg. Log Prob.	**3364.82**				
AIC	**6767.64**				

Table 29: Learned weights for constraint set D_{full} with targets pre-set to the mean (black), the mean plus 2 SDs (red), or the mean minus 2 SDs (blue) of the natural class they constrain. Weights of 0 are omitted for visual clarity.

In light of these results, we can at the very least conclude that it is not safe to assume that the mathematically optimal constraint targets lie near the means of the natural class of sounds that they constrain, and that this is due to the fact each output duration is the product of multiple, competing constraints. The remainder of the chapter will therefore focus on learning weights and targets simultaneously for various constraint sets—this turns out to produce grammars which achieve a much better fit and a higher degree of informativity.

6.5. Simultaneous weight and target learning

Learning constraint targets is a computationally more difficult endeavor than learning constraint weights alone. While the implementation of the gradient descent step is much the same—the optimization function gradually changes the weight and target parameters in a way

which improves model fit—it is no longer guaranteed to converge, or to converge at the globally optimal parameters, because the learning space is no longer guaranteed to be convex.

Several sets of constraints are trained on the data: first, the full set of DURATION constraints, D_{full}, then two subsets thereof, D_{short} and D_{long}, and finally several sets of STRETCH and SQUEEZE constraints: S_{full}, S_{sparse}, S_{short}, and S_{long}.

6.5.1. DURATION grammar fragments

6.5.1.1. Constraint Set D_{full}

The full set of DURATION constraints (set D_{full}) was trained on the data, learning both weights and targets. Four training runs were conducted. The results of all four runs are given in Table 30, and the results of the second run (which had the best AIC by a tiny margin) are rearranged in Table 31,where the natural class means are also provide for ease of comparison.

	Run 1		Run 2		Run 3		Run 4	
constraint	**weight**	**target**	**weight**	**target**	**weight**	**target**	**weight**	**target**
DUR vowel	0.01	1.58	0.05	2.28	0.00	1.46	0.00	1.71
DUR non-high	0.10	2.36	0.07	2.48	0.06	1.91	0.08	2.62
DUR low	0.28	4.38	0.27	4.33	0.27	4.38	0.05	1.66
DUR tense	0.32	5.94	0.31	6.16	0.30	6.50	0.42	5.55
DUR open	0.01	1.63	0.10	2.49	---	2.44	0.00	1.86
DUR singleton coda	0.72	1.77	0.28	2.30	0.70	1.72	0.52	1.45
DUR pre-voiced	0.21	1.75	0.22	3.65	0.25	1.92	0.44	1.95
DUR post b	0.32	4.73	0.32	4.96	0.36	3.00	0.36	4.82
DUR accented	0.11	3.23	0.04	2.68	0.16	5.11	0.13	3.57
DUR phrase-final	0.00	1.65	0.00	4.45	0.00	3.62	0.01	2.57
DUR high	0.25	-0.11	0.23	-0.30	0.22	-0.58	0.23	-0.22
DUR non-low	0.52	1.69	0.49	1.59	0.49	1.83	0.38	0.39
DUR lax	0.14	1.02	0.12	1.02	0.12	1.75	0.24	2.43
DUR closed	0.07	2.27	0.60	1.01	0.05	2.35	0.05	4.30
DUR complex coda	0.79	1.26	0.34	1.01	0.77	1.20	0.58	0.79
DUR pre-voiceless	2.08	0.92	2.09	1.13	2.12	0.96	2.30	1.04
DUR post m	0.19	0.56	0.18	0.78	0.22	-1.53	0.22	1.36
DUR unaccented	0.47	-0.02	0.41	-0.57	0.53	0.90	0.49	0.23
DUR phrase-medial	0.81	0.01	0.81	0.01	0.81	0.01	0.82	0.03
Neg. Log Prob.	**3321.66**		**3321.66**		**3321.66**		**3321.66**	
AIC	**6719.32**		**6719.32**		**6719.32**		**6719.32**	

Table 30: Learned weights and targets (in deciseconds) from four training runs for constraint set D_{full}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted—0.00 indicates a positive value smaller than 0.005.

constraint	weight	target	(mean)	constraint	weight	target	(mean)
DUR high	0.23	-0.30	1.50	DUR non-high	0.07	2.48	1.77
DUR non-low	0.49	1.59	1.60	DUR low	0.27	4.33	1.91
DUR lax	0.12	1.02	1.55	DUR tense	0.31	6.16	1.81
DUR closed	0.60	1.01	1.59	DUR open	0.10	2.49	2.16
DUR complex coda	0.34	1.01	1.39	DUR singleton coda	0.28	2.30	1.68
DUR pre-voiceless	2.09	1.13	1.42	DUR pre-voiced	0.22	3.65	1.93
DUR post m	0.18	0.78	1.51	DUR post b	0.32	4.96	1.79
DUR unaccented	0.41	-0.57	1.48	DUR accented	0.04	2.68	1.79
DUR phrase-medial	0.81	0.01	1.44	DUR phrase-final	0.00	4.45	1.89
DUR vowel	0.05	2.28	1.66				
Neg. Log Prob.	**3321.66**						
AIC	**6719.32**						

Table 31: Learned weights and targets (in deciseconds) from training run #2 of constraint set D_{full}, with mean durations of the class constrained for comparison. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). 0.00 indicates a positive value smaller than 0.005.

As might be expected, learning targets automatically results in models which are more informative than those in which targets were pre-selected, even adjusting for the fact that these models have double the number of parameters. A few interesting observations can be made about the learned targets. First, rather than being near the means for the classes of sounds they constrain, the constraints on relatively longer classes have targets longer than the mean, and the constraints on relatively shorter classes have targets shorter than the mean. In some cases, these targets are quite extreme, reaching 616 ms for DURATION[tense] and -57 ms for DURATION[unaccented].

Clearly, many of these learned "targets" cannot reasonably be actual phonetic targets in the traditional sense of being articulatory or acoustic goals for the speaker; after all, how would one attempt to produce a vowel with negative duration? However, this pattern of extreme targets will recur throughout this chapter, and even the best models found—models which turn out to fit the

data qualitatively very well—will make use of them. A discussion of why these targets are the ones learned, and how they should be interpreted, will be given at the end of the chapter.

A second observation is that the four training runs converged on different parameters values, but found models which were equally good. This kind of inconsistent convergence behavior is indicative of a learning space which is not convex, but instead has multiple local optima. These optima could even lie along large "valleys" or "ravines" in which certain sets of parameters can be simultaneously adjusted in a way which doesn't affect goodness of fit. Flemming and Cho (2017), who use gradient descent to learn both weights and targets for a phonetic harmonic grammar, encounter the same behavior: they similarly find multiple distinct local optima, but the parameter values at these optima produce quantitatively and qualitatively similar results. It seems likely that, at least for constraint sets involving targets, phonetic grammars may simply have more than one way of accounting for the data.

A final observation is that the model assigns less weight on average to the constraints on the relatively longer categories of sounds, and more to the relatively shorter categories. For example, DURATION[unaccented] is given a weight of 0.41, and DURATION[accented] a weight of only 0.04. This overall pattern—that shorter categories are more constrained—is in line with the "Synergistic Shortening" explanation offered for the Hyperadditive Lengthening Generalization in Section 5.5.1.2, which was that hyperadditive lengthening (really hypo-additive shortening) will always result when all or most duration effects are shortening effects (constraining phrase-medial vowels, and high vowels, and vowels in closed syllables, and so forth).

6.5.1.2. Constraint Sets D_{short} and D_{long}

This last observation calls into question whether constraints on longer categories are needed at all, or whether a shortening-only grammar would in fact be more informative. To test this, a constraint set consisting of DURATION[vowel] and the nine DURATION constraints on relatively shorter categories (set D_{short}) was trained on the data, learning both weights and targets. Four training runs were conducted. The results of all four runs are given in Table 32, and the results of the first run, which had the best AIC, are rearranged in Table 33.

	Run 1		**Run 2**		**Run 3**		**Run 4**	
constraint	**weight**	**target**	**weight**	**target**	**weight**	**target**	**weight**	**target**
DUR vowel	1.25	5.49	1.18	5.69	1.33	5.24	1.25	5.43
DUR high	0.07	-5.92	0.09	-4.11	0.04	-10.37	0.03	-18.51
DUR non-low	0.11	-18.83	0.17	-11.00	0.06	-31.84	0.11	-17.65
DUR lax	0.05	-28.09	0.12	-10.58	0.05	-25.21	0.06	-24.78
DUR closed	0.92	1.79	0.93	1.81	0.88	1.79	0.91	1.81
DUR complex coda	0.03	-11.65	0.12	-1.72	0.01	-37.38	0.04	-7.56
DUR pre-voiceless	1.87	0.82	1.78	0.78	1.86	0.82	1.81	0.80
DUR post m	0.04	-25.06	0.08	-13.77	0.03	-32.77	0.04	-29.25
DUR unaccented	0.42	-0.58	0.27	-1.92	0.34	-1.13	0.41	-0.65
DUR phrase-medial	0.75	-0.16	0.78	-0.06	0.77	-0.10	0.79	-0.04
Neg. Log Prob.	**3323.65**		**3324.32**		**3323.66**		**3323.75**	
AIC	**6687.29**		**6688.64**		**6687.32**		**6687.50**	

Table 32: Learned weights and targets (in deciseconds) from four training runs on constraint set D_{short}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long).

constraint	weight	target	(mean)
DUR high	0.07	-5.92	1.50
DUR non-low	0.11	-18.83	1.60
DUR lax	0.05	-28.09	1.55
DUR closed	0.92	1.79	1.59
DUR complex coda	0.03	-11.65	1.39
DUR pre-voiceless	1.87	0.82	1.42
DUR post m	0.04	-25.06	1.51
DUR unaccented	0.42	-0.58	1.48
DUR phrase-medial	0.75	-0.16	1.44
DUR vowel	1.25	5.49	1.66
Neg. Log Prob.	**3323.65**		
AIC	**6687.29**		

Table 33: Learned weights and targets (in deciseconds) from training run #1 on the constraint set D_{short}, with mean durations of the class constrained for comparison. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long).

This simpler "shortening only" model achieves almost as good a fit to the data as does the full model, using a much smaller number of parameters (20 instead of 38), and, as a result, is a more informative model by AIC.

The learned target values are even more extreme than before, ranging from the target of DURATION[vowel] at half a second, and that of DURATION[lax] at a whopping -2.8 seconds.

For completeness, and to show that this improvement in informativity is due to focusing on shortening constraints in particular, and not just removing half of the parameters arbitrarily, a constraint set consisting of DURATION[vowel] and the nine DURATION constraints on relatively longer categories (set D_{long}) was trained on the data. The resulting models, in addition to being inconsistent between training runs, are both qualitatively and quantitatively awful (Table 34).

	Run 1		Run 2		Run 3		Run 4	
constraint	**weight**	**target**	**weight**	**target**	**weight**	**target**	**weight**	**target**
DUR vowel	2.65	0.39	2.57	0.42	1.82	0.22	2.39	0.27
DUR non-high	0.01	50.00[49]	0.04	13.81	0.01	50.00	0.03	15.75
DUR low	0.01	50.00	0.01	49.99	0.00	50.00	0.12	5.77
DUR tense	0.02	50.00	0.02	47.95	0.03	50.00	0.03	33.34
DUR open	0.04	50.00	0.11	17.93	---	1.32	0.04	50.00
DUR singleton coda	0.00	50.00	0.00	-50.00	0.78	1.27	0.31	2.16
DUR pre-voiced	0.03	50.00	0.03	50.00	0.02	50.00	0.04	33.26
DUR post b	0.02	50.00	0.02	50.00	0.02	50.00	0.06	18.46
DUR accented	0.02	50.00	0.04	20.97	0.01	50.00	0.02	35.22
DUR phrase-final	0.02	50.00	0.02	50.00	0.02	50.00	0.02	50.00
Neg. Log Prob.		**3482.75**		**3492.16**		**3573.19**		**3483.53**
AIC		**7005.51**		**7024.32**		**7186.38**		**7007.05**

Table 34: Learned weights and targets (in deciseconds) from four training runs, using constraint set D_{long}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted—0.00 indicates a positive value smaller than 0.005.

6.5.2. STRETCH and SQUEEZE grammar fragments

In Chapter 4, a generalization of the DURATION constraint family was proposed, splitting this constraint into STRETCH and SQUEEZE constraints, identical to DURATION constraints except that they do not penalize candidates which lie to one side of their target. As shown in that chapter, these asymmetrical constraints allow for the possibility of non-normality in the predicted distributions, and variation, for example, in these distributions' kurtoses. Such variation was seen to be empirically present in the experimental results in Chapter 5.

[49] Recall that 50.00 ds (5 s) was the upper bound on targets given to the optimization function.

6.5.2.1. Constraint Set S_{full}

First, a constraint set analogous to D_{full}, but with each DURATION constraint replaced with one STRETCH and one SQUEEZE constraint, was trained on the data (set S_{full}, 'S' standing for both 'STRETCH' and 'SQUEEZE'). The results are shown in Table 35.

	Run 1		**Run 2**		**Run 3**		**Run 4**	
constraint	**weight**	**target**	**weight**	**target**	**weight**	**target**	**weight**	**target**
SQUEEZE vowel	---	2.01	---	1.41	---	1.66	---	1.81
SQUEEZE non-high	---	1.37	---	1.64	---	1.66	0.03	0.67
SQUEEZE low	0.05	1.66	0.05	1.88	0.01	1.52	0.17	1.76
SQUEEZE tense	---	1.86	0.18	1.15	---	1.63	---	1.82
SQUEEZE open	---	1.51	---	1.73	---	1.89	---	1.75
SQUEEZE singleton coda	0.28	1.45	0.17	-0.06	0.45	1.40	0.14	1.56
SQUEEZE pre-voiced	0.01	0.99	0.01	0.95	---	1.93	0.12	-0.68
SQUEEZE post b	---	1.84	---	1.69	---	1.68	---	1.75
SQUEEZE accented	0.16	1.03	0.06	1.37	0.23	1.48	0.24	1.61
SQUEEZE phrase-final	0.01	1.55	0.24	1.59	---	1.44	---	1.95
SQUEEZE high	0.11	-2.25	0.10	-2.77	0.13	-1.56	0.09	-3.99
SQUEEZE non-low	0.08	-0.23	---	-0.49	0.07	-0.35	---	1.70
SQUEEZE lax	---	1.71	---	1.65	---	1.84	---	1.54
SQUEEZE closed	0.17	1.25	0.09	1.24	0.01	1.67	0.03	1.20
SQUEEZE complex coda	0.23	0.04	0.19	-0.17	0.27	-0.36	0.27	1.44
SQUEEZE pre-voiceless	0.74	-0.83	0.80	-0.50	0.93	-0.14	0.88	-0.66
SQUEEZE post m	0.12	-1.44	0.13	-1.61	0.14	-0.84	0.10	-3.50
SQUEEZE unaccented	0.21	-3.41	0.16	-4.27	0.24	-2.76	0.16	-4.89
SQUEEZE phrase-medial	0.24	-3.65	0.22	-5.34	0.23	-3.92	0.17	-5.76

continued on next page…

…continued

constraint	weight	target	weight	target	weight	target	weight	target
STRETCH vowel	---	1.06	3.02	0.71	0.94	0.63	3.98	0.65
STRETCH non-high	0.65	1.57	0.19	2.08	0.58	1.63	0.32	1.50
STRETCH low	0.34	3.46	0.30	4.21	0.42	3.20	0.17	6.36
STRETCH tense	0.52	4.68	0.40	5.85	0.49	4.84	0.60	4.27
STRETCH open	0.60	1.29	1.97	1.35	0.72	1.30	1.74	1.21
STRETCH singleton coda	0.78	2.12	1.42	2.10	1.16	0.67	1.15	2.13
STRETCH pre-voiced	0.00	0.54	---	0.80	0.77	2.00	0.00	0.59
STRETCH post b	0.64	3.10	0.83	2.70	0.71	2.95	0.75	2.70
STRETCH accented	0.03	1.98	0.12	2.12	0.03	2.10	0.01	0.56
STRETCH phrase-final	0.80	2.10	---	0.91	0.48	2.26	0.96	2.07
STRETCH high	1.89	0.74	---	1.39	---	1.19	---	1.17
STRETCH non-low	0.81	1.28	1.34	1.35	1.50	1.36	---	1.62
STRETCH lax	0.76	0.61	---	1.24	0.52	0.60	---	1.21
STRETCH closed	1.99	0.74	---	3.12	1.60	0.75	1.97	0.68
STRETCH complex coda	0.74	1.98	1.83	1.70	---	1.33	1.41	1.67
STRETCH pre-voiceless	0.41	2.41	0.48	2.02	1.03	2.09	0.49	2.01
STRETCH post m	0.42	0.13	0.39	0.46	1.03	0.57	1.49	0.60
STRETCH unaccented	1.08	0.92	1.61	0.88	1.22	0.78	2.10	0.62
STRETCH phrase-medial	2.53	1.46	1.10	1.42	1.91	1.44	3.46	1.40
Neg. Log Prob.		**3293.19**		**3293.93**		**3293.82**		**3291.77**
AIC		**6738.39**		**6739.87**		**6739.63**		**6735.54**

Table 35 : Learned weights and targets (in deciseconds) from four training runs, using constraint set S_{full}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted—0.00 indicates a positive value smaller than 0.005.

While the learned grammars with the constraint set S_{full} achieve a better fit to the data than do those with D_{full}, they do so only because they have many more parameters (a whopping 76, not that far off from 128, the number of experimental conditions), and as a result have a worse AIC.

However, as before, it is easy to imagine that employing this full set of constraints might be extraneous. In particular, it may not be necessary to provide STRETCH constraints on short categories like phrase-medial, or, conversely, SQUEEZE constraints on longer categories like phrase-final.

6.5.2.2. Constraint Set S_{sparse}

A second set of constraints consisting of STRETCH[vowel], SQUEEZE[vowel], STRETCH constraints on the nine relatively longer categories, and SQUEEZE constraints on the nine relatively shorter categories (set S_{sparse}) was trained on the data. The results of all four runs are given in Table 36 and the results of the second run, which had the best AIC, are rearranged in Table 37.

	Run 1		Run 2		Run 3		Run 4	
constraint	**weight**	**target**	**weight**	**target**	**weight**	**target**	**weight**	**target**
STRETCH vowel	4.38	1.31	4.54	1.29	4.44	1.33	4.19	1.37
STRETCH non-high	1.06	1.29	1.08	1.28	0.50	1.54	0.33	1.91
STRETCH low	0.11	6.31	0.12	7.62	0.13	7.46	0.26	3.38
STRETCH tense	0.81	3.51	0.85	3.48	0.89	3.34	0.73	3.73
STRETCH open	2.42	1.08	0.41	0.91	0.34	0.30	0.70	1.05
STRETCH singleton coda	5.05	0.73	8.21	0.70	0.53	2.14	0.53	2.14
STRETCH pre-voiced	0.80	2.13	0.74	2.17	0.22	0.54	---	1.10
STRETCH post b	0.65	3.20	1.10	2.20	1.19	2.17	0.70	3.13
STRETCH accented	0.19	2.01	0.11	2.30	0.13	4.25	0.22	2.11
STRETCH phrase-final	0.02	4.01	0.04	6.24	0.26	0.25	---	2.68
SQUEEZE vowel	---	1.68	---	3.60	0.35	2.84	---	1.92
SQUEEZE high	0.08	-3.93	0.08	-4.13	0.08	-3.66	0.10	-2.46
SQUEEZE non-low	0.06	-2.02	0.40	2.81	0.00	0.94	0.07	-2.47
SQUEEZE lax	0.20	1.80	0.41	2.76	0.14	1.66	0.04	1.70
SQUEEZE closed	0.41	1.22	0.48	1.32	0.48	1.42	0.49	1.44
SQUEEZE complex coda	0.10	-2.29	0.13	-1.70	0.29	1.20	0.32	1.25
SQUEEZE pre-voiceless	2.03	1.15	1.94	1.14	1.16	0.42	1.22	0.47
SQUEEZE post m	0.07	-2.22	0.11	-4.09	0.13	-3.16	0.10	-1.19
SQUEEZE unaccented	0.11	-6.12	0.11	-6.64	0.10	-5.17	0.14	-4.47
SQUEEZE phrase-medial	0.20	-5.03	0.17	-5.54	0.16	-6.74	0.21	-4.89
Neg. Log Prob.	**3301.05**		**3299.67**		**3300.53**		**3302.85**	
AIC	**6682.11**		**6679.34**		**6681.06**		**6685.69**	

Table 36: Learned weights and targets (in deciseconds) from four training runs, using constraint set S_{sparse}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted—0.00 indicates a positive value smaller than 0.005.

constraint	weight	target	(mean)	constraint	weight	target	(mean)
SQ high	0.08	-4.13	1.50	STR non-high	1.08	1.28	1.77
SQ non-low	0.40	2.81	1.60	STR low	0.12	7.62	1.91
SQ lax	0.41	2.76	1.55	STR tense	0.85	3.48	1.81
SQ closed	0.48	1.32	1.59	STR open	0.41	0.91	2.16
SQ complex coda	0.13	-1.70	1.39	STR singleton coda	8.21	0.70	1.68
SQ pre-voiceless	1.94	1.14	1.42	STR pre-voiced	0.74	2.17	1.93
SQ post m	0.11	-4.09	1.51	STR post b	1.10	2.20	1.79
SQ unaccented	0.11	-6.64	1.48	STR accented	0.11	2.30	1.79
SQ phrase-medial	0.17	-5.54	1.44	STR phrase-final	0.04	6.24	1.89
SQ vowel	---	3.60	1.66	STR vowel	4.54	1.29	1.66
Neg. Log Prob.	**3299.67**						
AIC	**6679.34**						

Table 37: Learned weights and targets (in deciseconds) from training run #2 on the constraint set S_{sparse}, with mean durations of the class constrained for comparison. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted.

The S_{sparse} constraint set turned out to produce the most informative models, out of all of those tried, surpassing D_{short} by a large margin.

Targets here seem to have been employed in two very different ways here by the learner. The first is analogous to the way they are used by the D_{full} model and especially the D_{short} model: a rather extreme target is paired with a small weight, which results in a lengthening or shortening effect that applies over the whole range of duration candidates—this is true for constraints SQEEZE[high], STRETCH[low], and many others.[50]

[50] It is interesting to note that the "flat" portion of each of these constraints lies way out in an extreme duration region, much shorter or longer than any candidate which is being given any significant probability. This means that if these constraints had been the symmetrical DURATION constraints instead of STRETCH or SQUEEZE, it would make no difference; the fact that they are asymmetrical has no consequence in these cases.

However, other constraints have targets on the opposite end of the duration spectrum, where they are active over only either the smallest or largest parts of the duration range. For example, STRETCH[singleton coda] has a very high weight, but only penalizes candidates shorter than 70 ms. This constraint is acting as a kind of soft "minimum duration" or "maximum compressibility" constraint, serving only to rule out very short vowels (at least in singly-closed syllables)—and with good reason: in the training data, only 14 of the 736 singleton tokens in this natural class have durations shorter than 70 ms. On the shortening side, SQEEZE[pre-voiceless] applies only to candidates with duration greater than 114 ms: while its weight is not high enough to create a ceiling effect, this constraint nevertheless only effects part of the positive duration range, and is therefore being used in a different way than constraints like SQUEEZE[high], which here penalize all positive duration candidates.

6.5.2.3. Constraint Sets S_{short} and S_{long}

Since using a "shortening only" grammar resulted in an improvement in informativity for the DURATION constraint family, the same could *a priori* be true for STRETCH and SQUEEZE grammars. A set of constraints consisting of STRETCH and SQUEEZE[vowel] along with SQUEEZE constraints on the nine relatively shorter categories (set S_{short}) was trained on the data, and the results of all four runs are given in Table 38.

As before, for the sake of symmetry, a "lengthening only" constraint set consisting of STRETCH and SQUEEZE[vowel] along with STRETCH constraints on the nine relatively longer categories (set S_{long}) was also trained on the data, and the results of all four runs are given in Table 39.

	Run 1		Run 2		Run 3		Run 4	
constraint	**weight**	**target**	**weight**	**target**	**weight**	**target**	**weight**	**target**
STRETCH vowel	1.57	3.19	1.39	5.23	1.53	4.91	1.43	5.14
SQUEEZE vowel	0.00	0.95	0.00	2.04	0.00	1.47	0.00	2.78
SQUEEZE high	0.04	-20.22	0.16	-1.73	0.15	-1.86	0.16	-1.78
SQUEEZE non-low	0.32	2.86	0.16	-11.45	0.09	-21.60	0.15	-12.81
SQUEEZE lax	0.02	-28.67	0.34	-2.39	0.13	-8.88	0.28	-3.20
SQUEEZE closed	0.90	2.81	0.43	1.13	0.47	1.11	0.44	1.12
SQUEEZE complex coda	1.96	4.27	0.27	0.25	0.24	0.13	0.30	0.35
SQUEEZE pre-voiceless	1.56	0.75	1.63	0.75	1.67	0.76	1.67	0.77
SQUEEZE post m	0.03	-25.37	0.10	-10.01	0.04	-26.76	0.06	-18.77
SQUEEZE unaccented	0.39	-0.34	0.45	-0.44	0.46	-0.43	0.44	-0.52
SQUEEZE phrase-medial	0.77	0.22	0.91	0.18	0.92	0.17	0.90	0.15
Neg. Log Prob.	**3459.68**		**3334.78**		**3333.91**		**3334.08**	
AIC	**6963.36**		**6713.56**		**6711.81**		**6712.17**	

Table 38: Learned weights and targets (in deciseconds) from four training runs, using constraint set S_{short}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted—0.00 indicates a positive value smaller than 0.005.

	Run 1		Run 2		Run 3		Run 4	
constraint	**weight**	**target**	**weight**	**target**	**weight**	**target**	**weight**	**target**
STRETCH vowel	8.00	1.17	4.60	1.62	5.94	1.38	5.43	1.61
STRETCH non-high	0.78	2.12	1.20	1.78	0.73	2.18	0.08	8.96
STRETCH low	0.08	11.38	0.18	5.92	0.20	5.79	0.04	14.31
STRETCH tense	2.22	2.08	0.44	4.87	1.39	2.46	0.77	3.50
STRETCH open	0.12	18.62	0.07	27.66	0.11	19.60	0.08	27.21
STRETCH singleton coda	0.45	2.12	5.71	0.54	1.72	0.18	8.42	0.70
STRETCH pre-voiced	0.11	14.05	0.08	20.00	0.10	16.18	0.11	16.01
STRETCH post b	0.82	3.10	0.74	3.31	0.63	3.50	0.76	3.26
STRETCH accented	0.12	9.49	0.13	8.61	0.05	20.47	0.05	21.20
STRETCH phrase-final	0.08	18.96	0.06	22.52	0.09	15.93	0.08	19.88
SQUEEZE vowel	0.20	-18.05	0.19	-23.42	0.27	-13.84	0.16	-30.68
Neg. Log Prob.	**3349.05**		**3347.93**		**3350.92**		**3341.64**	
AIC	**6742.10**		**6739.85**		**6745.84**		**6727.28**	

Table 39: Learned weights and targets (in deciseconds) from four training runs, using constraint set S_{long}. Relative weight is shown in yellow shading, target values on a scale from blue (short) to red (long). Weights of 0 are omitted—0.00 indicates a positive value smaller than 0.005.

In fact, neither of these models performed well, showing inconsistent convergence results with poor fits to the data, suggesting that both some STRETCH and some SQUEEZE constraints were in fact necessary conditions for the success of the S_{sparse} grammar.

Table 40 summarizes the results of all grammars for which weights and targets were learned.

Constraint Set	**Con-straints**	**Lowest AIC**	**Highest AIC**	**Notes**
D_{full}	19	6719.32	6719.32	Viable models
D_{short}	10	6687.29	6688.64	Viable models (second best)
D_{long}	10	7005.51	7186.38 !	Inconsistant convergence
S_{full}	38	6735.54	6739.87	Too many parameters, likely overfitting
S_{sparse}	20	6679.34	6685.69	Viable models (best)
S_{short}	11	6711.81	6963.36 !	Inconsistant convergence
S_{long}	11	6727.28	6745.84 !	Inconsistant convergence

Table 40: Model performance for the constraint sets learned.

6.6. Predictions of learned models

6.6.1. Predictions of sample means

The S_{sparse} grammar was able to fairly accurately predict the location of the mean durations of the 128 experimental conditions (Figure 37), accounting for 90% (R^2 = .901) of their variation. Encouragingly, there doesn't seem to be any pattern in the residual, meaning that the grammar is predicting the means of relatively longer and relatively shorter conditions equally well.

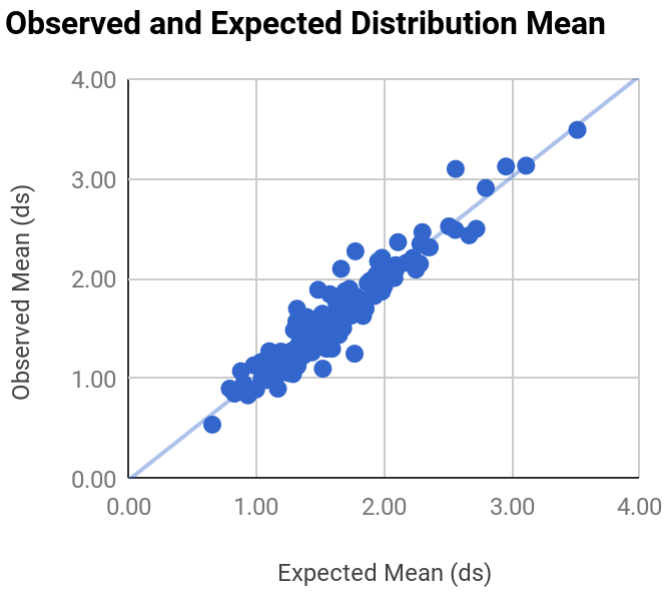

Figure 37: The means of the duration distributions predicted by the S_{sparse} grammar compared with observed sample means from the 128 experimental conditions.

However, focusing on means provides only a very limited picture. Unlike most models, maxent grammars do not just predict means for each input to the grammar, but entire probably distributions: we can compare these directly to histograms of the training data. Six such comparisons, with a range of different targets and prosodic conditions, are given in Figure 38.

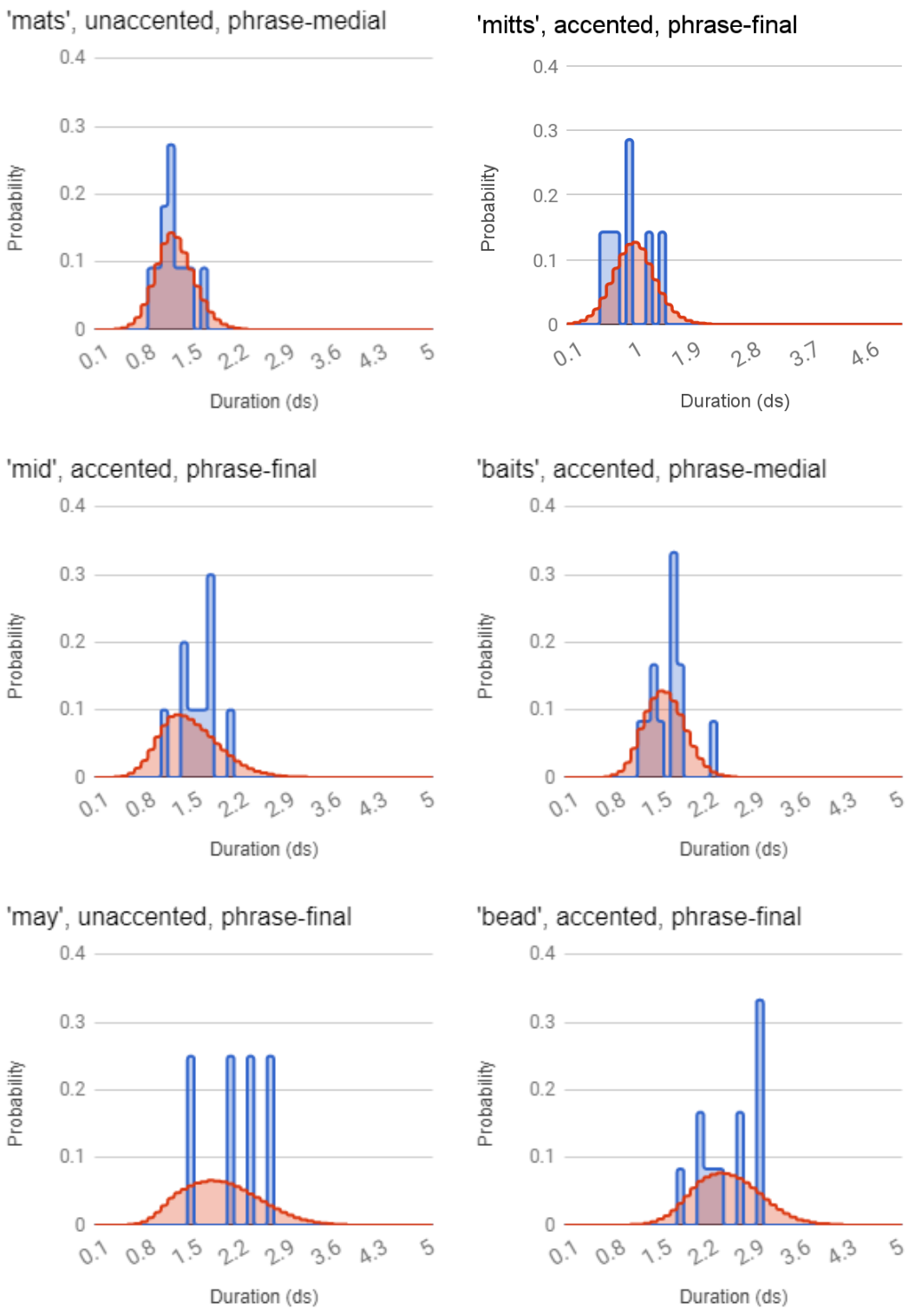

Figure 38: Predicted distributions of the S_{sparse} grammar (red), and observed histograms from the training data (blue), for six of the 128 experimental conditions.

6.6.2. Predictions of standard deviations and kurtoses

As can be observed in Figure 38, the S_{sparse} grammar not only predicts that there will be differences in the means of the 128 conditions, but also differences in the shapes of their distributions.

For example, the grammar predicts that there will be variation in these distributions' standard deviations, and this predicted variation is correlated with variation in standard deviation seen in the training data (Figure 39), accounting for 45% (R^2 = .450) of the observed variation.

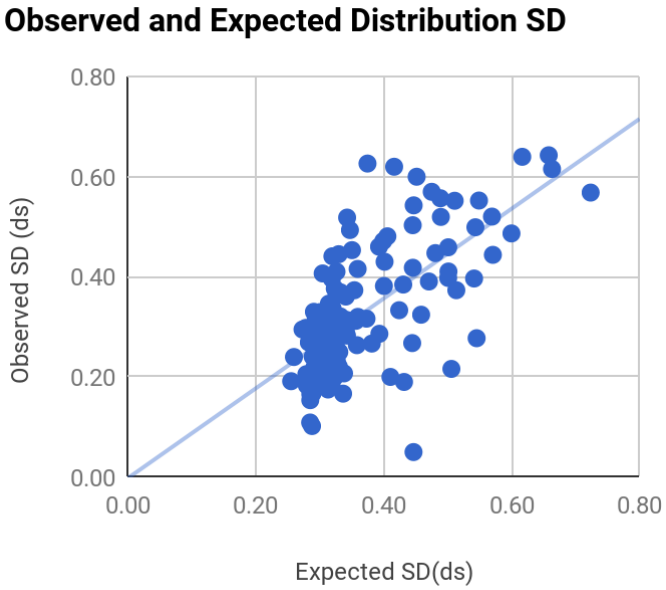

Figure 39: The standard deviations of the duration distributions predicted by the S_{sparse} grammar compared with observed sample standard deviations from the 128 experimental conditions.

The grammar also predicts that there will be some variation in these distributions' kurtoses, which is only possible due to the asymmetrical nature of STRETCH and SQUEEZE constraints. However, predicted kurtoses do not match those seen in the data very well (Figure 40), accounting for only 1% (R^2 = .013) of the variation. In fact, all of the predicted values are positive, ranging from 0.04 to 0.62, while the observed ones range from -1.53 to 1.88.

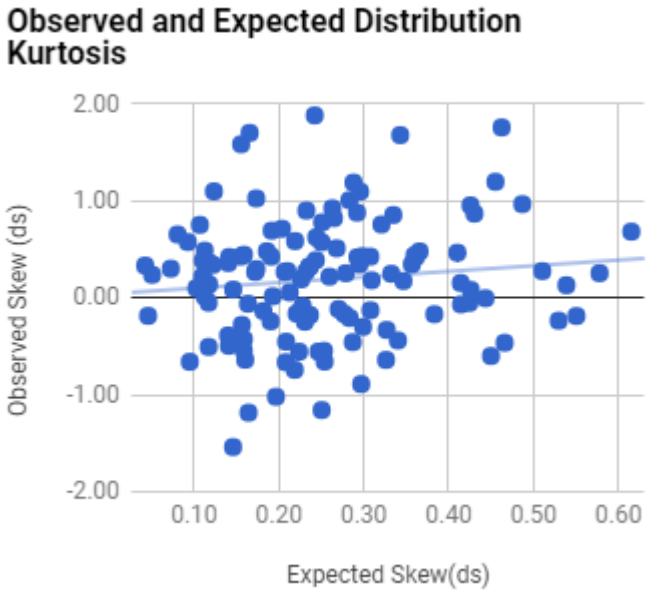

Figure 40: The kurtoses predicted by the S_{sparse} grammar compared with observed sample kurtoses from the 128 experimental conditions.

This may in part be because many of the samples were quite small, and accurately estimating the skewness of a population requires a fairly large sample. In other words, a good deal of the variation in kurtosis in the observed distributions was probably due to noise.

6.6.3. Conditioned and unconditioned variation

Empirical evidence was found (section 5.4.7) that, when comparing sets experimental conditions which differed only in one way (say, phrase-final vs. phrase-medial), the category of conditions which showed more phonologically conditioned variation (as estimated by the standard distribution of the sample means) was generally also the category which showed more random variation (as estimated by the mean of the sample standard deviations), and these two kinds of variation were correlated (Table 22, Figure 35).

The distributions generated by the S_{sparse} grammar also showed this correlation ($R^2 = 0.84$) (Table 41, Figure 41). This is expected, since the degrees of both types of variation are determined by the weights of the very same constraints (section 4.1.5.1).

Feature / Comparison	Duration	
	Δ Uncond. Var.	Δ Cond. Var.
high (high vs. mid)	0.056	0.218
low (ɛ vs. æ)	-0.031	-0.132
tense (ɛ, ɪ vs. eɪ, i)	0.024	-0.014
nasal onset (m vs. b)	0.010	0.063
closed (t vs. Ø)	-0.010	-0.124
voiceless coda (t vs. d)	0.000	-0.038
complex coda (ts vs. t)	-0.009	-0.048
accented	0.003	0.008
phrase-final	-0.003	-0.041

Table 41 : Predicted differences between longer and shorter categories, across a number of phonological features, in their propensity toward conditioned variation, and unconditioned variation, for both duration (left) and log-duration (right), according to the S_{sparse} grammar.

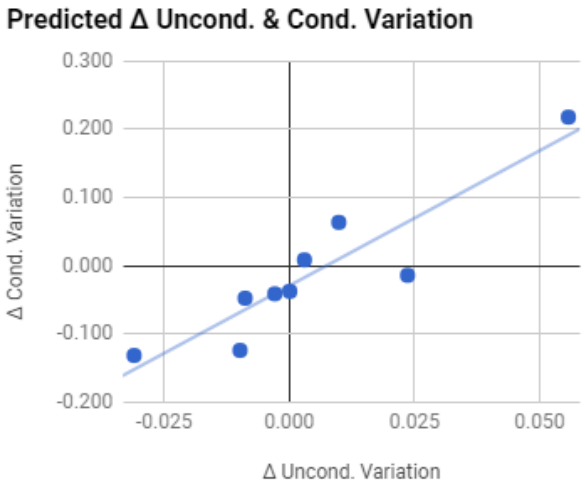

Figure 41: Predicted differences between longer and shorter categories, across each of the nine binary phonological features, in their propensity to show conditioned and unconditioned variation, for both duration (left) and log-duration (right), according to the S_{sparse} grammar.

6.7. Discussion

Learning the parameters of maxent grammars with phonetic constraints appears to be feasible. This in and of itself is a novel and encouraging result, opening the door to future work on maxent phonetic grammars.

With respect to modeling duration in particular, some of the constraint sets produce models which seem to fit the data well, as long as the learner is allowed to learn both the weights and the targets of the constraints. Of the grammars which used only the DURATION constraint family, models which contained only shortening effects (D_{short}) were more informative than models with only lengthening effects (D_{long}) and models with both shortening and lengthening effects (D_{full}), and in the latter, shortening constraints are given more weight. There are a number of reasons that this is the case. In Chapter 5, it was shown that a bias towards shortening constraints would automatically derive the hyperadditive lengthening generalization, due to synergistic shortening (section 5.5.1.2). Constraining only or primarily the relatively shorter categories also allows the grammar to derive another pattern seen in the data: that relatively shorter categories have relatively narrower distributions (section 5.4.7).

Even better models were achieved by employing the asymmetrical STRETCH and SQUEEZE constraints introduced in Chapter 4, specifically with the constraint set (S_{sparse}), which contained STRETCH constraints for natural classes which are comparatively long, and vice versa for SQUEEZE, as well as a global STRETCH and a SQUEEZE constraint, but omitted SQUEEZE constraints on longer classes and STRETCH constraints on shorter ones (leaving out constraints which would have effects like phrase-final shortening or phrase-medial lengthening: perhaps an intuitive omission). The fact that this constraint set (S_{sparse}) fit the data better than a similarly sized constraint set of symmetrical constraints (D_{full}) is of interest, since only the former is able to capture

distributions which are not Gaussian. Some evidence from Chapter 5 pointed towards skewness being a non-arbitrary feature of the different samples in the experimental data, and therefore to non-normality in free variation for phonetic variables like duration. While the learned model in this case didn't do a good job of capturing the observed variation is kurtosis (possibly due to there being too much noise), the existence of such variation, if it can be verified, is reason to suspect that asymmetrical constraints should have a place in phonetic grammars.

While the learned models produce fairly good results, they do so by learning constraint targets which are extremely counter-intuitive, seriously calling into question the explanatory adequacy of these grammars. Negative durations, durations of over half a second, and so on are clearly not realistic values for phonetic targets in the traditional sense. Why the learner chooses these values, and what this means, deserves some explanation.

6.7.1. Target learning: a post-mortem

In prior work by linguists using targeted phonetic constraints, and harmonic grammars thereof, constraint targets are clearly thought of as representing the articulatory or acoustic goals of the speaker with respect to some sound or phonological constituent. For example, the proof of concept given by Flemming (2001) relating to CV coarticulation (reproduced here in section 3.3.1) involves a constraint IDENT(V) with a target representing the default F2 of some vowel, a constraint IDENT(C) with a target representing the default F2 locus of some consonant, and a constraint MINEFFORT penalizing F2 transitions in CV sequences. In the duration domain, Katz (2010) constructs examples of harmonic grammars consisting of DURATION constraints with targets which look reasonably like empirically observed duration averages for the kinds of syllables and segments they constrain, and Braver (2013) and Hayes and Schuh (MS) do the same. All manage to use these constraints to create grammar fragments which, with some adjustment, do a good job

of predicting qualitative patterns seen in the average durations of segments or syllables under investigation.

Why, then, did the maxent learner developed here, when presented with duration data for front vowels in various phonological contexts, and with a set of constraints governing the duration of various natural classes thereof, fail to learn targets which were anywhere near the mean durations for these natural classes, instead learning apparently nonsensical ones that nevertheless yield grammars which predict the data well? Why, when forced to use these more reasonable targets (as in section 6.4), was the learner unable to learn weights which produced a quantitatively or qualitatively good fit to the data? There are two reasons for this.

6.7.1.1. Target ganging in complex grammars

The first reason relates to the complexity of the grammar, and in particular the highly orthogonal nature of the factors effecting front vowel duration, as compared to those which were modeled in earlier works. In Flemming's coarticulation illustration, for example, if a vowel phoneme were to be produced in isolation, its F2 would be governed only by a single constraint: IDENT(V). In order to make the correct predictions, then, this constraint would need to use a target which represents the empirically observed realization of that vowel phoneme in isolation.

The situation in grammars like D_{full} and S_{sparse} is very different. Because so many factors are always at play in determining duration, any given input to the grammar will always be subject to multiple constraints, by virtue of being a vowel which has certain segmental properties, occurs in a certain segmental context, and occurs in a certain prosodic context, all of which have effects on duration that are enforced by constraints. Because of this, none of the constraints are the sole constraint active in any one tableau. The articulatory goals of the speaker for any sound or category

of sounds, rather than being determined by the target of a single constraint, are instead a function of the *combined* violation profiles of several constraints in tandem.

For example, in the D_{short} grammar, the constraints DUR[vowel] (w=1.25, t=5.49 ds) and DUR[lax] (w=0.05, t=−28.09 ds) both seem to have extremely unrealistic phonetic targets. However, all the lax vowels in the data are *always* subject to both of these constraints, as well as to DUR[closed] (w=0.92, t=1.79 ds) for phonotactic reasons. While the violation functions of these individual constraint individually are parabolas centered on unrealistic targets, the sum of the three violation functions is a parabola centered around 3.20 ds (320 ms), the weighted average of the three constraints' targets. This number is a very reasonable articulatory target for what this grammar treats as the least marked lax vowel: /æ/ in the context /b_d/ in accented phrase-final position. (The experimentally observed mean is 313 ms.)

The takeaway here is that, in models where it is always the case that many constraints are active at once, the concrete phonetic goals of the speaker will not generally correspond to targets of any one constraint, but instead to weighted averages of the targets of all of the constraints which govern the sounds being produced.

6.7.1.2. Maxent grammars: slaves to variation

The second reason is that maxent grammars are different from harmonic grammars. While harmonic grammars predict a single winning candidate for each phonological input, while maxent grammars predict probability distributions over candidates. In the maxent learning step, the learner is not trying to learn a grammar which merely makes the correct predictions about category means, but instead to learn a grammar with predicted distributions which closely matched the distributions in the training data, in order to maximize its likelihood. In particular, if a learned grammar makes

the correct predictions about means, but predicts very narrow distributions when in fact lots of variation is present in the training data, it will be treated as fitting the data poorly. With that in mind, consider that in the phonetic maxent framework it is the absolute weights of the constraints which determine the standard deviations of the predicted distributions: large weights create highly constrained distributions with small standard deviations, and vice versa for low constraints (section 4.1.3). Therefore, the learner had to pick particular weights—relatively small ones, in fact—to accurately model the fact that the training data contained quite a lot of free variation.

However, the learner also wants to predict effect sizes correctly: if phrase-final vowels are very much longer than phrase-medial ones, for example, then the constraints on these categories had better enforce that distinction. However, the weights of these constraints, for reasons just discussed, cannot be made too large. The learner is left with a conundrum: how to create a large phonological effect size without making the predicted distributions too narrow. The solution, of course, is to make the targets more extreme! This result is a constraint which provides a weak pull towards a distant target. This is a viable solution because, again, the actual phonetic goals of the speaker are weighted averages of the targets of many constraints.

If the conditions in the training data had smaller standard deviations, the constraint weights would be larger, and the targets learned would be generally less extreme. Part of the reason for the small weights (and therefore extreme targets), then, might stem from the fact that I have used the grammar itself to model all of variance in the training data, when in fact much of this variance is probably not due to the grammar: the data were collected from multiple speakers, and these speakers varied in their speech rate, which was not modeled at all here, and instead treated as just a kind of free variation. Even if speech rate and every other linguistic variable were carefully controlled and modeled, though, it would still be difficult to determine how much of the variance

in the data should be attributed to the grammar. Since the distributions observed for different categories of sounds seem to differ, and to perhaps to be linguistically governed, the grammar probably has some role to play in governing variation—we need not abandon the maxent approach to phonetics simply because phonetic data is noisy—but at the same time a more plausible maxent phonetic grammar should probably output a distribution which has less variance than is actually observed. How to incorporate this insight into the maxent learning processes is a problem which is waiting to be solved.

6.7.2. Variation between learned grammars

When only constraint weights are learned, maxent grammars (including phonetic ones) have the attractive property that their parameter space is convex. This means that learning is guaranteed always to converge on the same set of weights, and these weights are guaranteed to be optimal. When both the targets and weights of phonetic constraints are treated as model parameters, this is no longer the case, and convergence behavior can be inconsistent, resulting in different sets of learned weights and targets each time. This kind of behavior is also reported by Flemming and Cho (2017) for weight and target learning using the HG framework.

However, this inconsistency is not necessarily cause for alarm. For most of the constraint sets the parameters learned on different training runs, while variable, were not qualitatively very different, and also resulted in overall grammars with similar fits to the data. What this means is that, at least for some constraints sets and some data, there are simply multiple phonetic grammars—often similar ones—which get the job done. Perhaps this is not even a fluke, but a feature, considering the fact that human learners do not always acquire identical grammars, even when exposed to similar linguistic input.

7. Conclusions

7.1. The long and the short of it[51]

The overarching empirical question of this dissertation was how the many phonological factors which affect phonetic durations interact with each other. Existing models treat the effect of each such factor as a coefficient, or ratio, which the default duration for the sound in question is multiplied by when that factor is present. When more than one effect applies, they are predicted to stack in a multiplicative way: this is equivalent to treating each factor as a fixed effect in a log-linear model.

However, throughout the literature on phonetic duration there are reports of significant interaction effects, where the duration which results when multiple factors are present is not what would be predicted by such a model (several of these interactions are replicated in Chapter 5), suggesting that the right generative model for phonetic duration may not be one in which a default duration is multiplied by a number of coefficients, but instead take some other shape.

7.1.1. Hyperadditive lengthening

The interactions reported are far from random. Instead, with few exceptions, they result in patterns where the longest cases are longer than would be expected, and the shortest cases are not as short as would be expected by a simple multiplicative model: this pattern is termed the Hyperadditive Lengthening Generalization (section 2.6). There are various possible theoretical explanations for this pattern (section 5.5.1), but its existence is interesting in itself, especially if it

[51] Duration-themed pun taken from Minkova, D. (2017, June). Personal interview.

turns out upon further investigation to be robust across even more pairs of factors, or if similar patterns can be found in phonetic variables other than duration.

7.2. Maxent phonetics

Maxent phonetic grammars, already in common use by phonologists, are potentially viable models for the realization of phonetic variables like duration. Like their phonological counterparts, they are able to predict variation over the possible realizations of these variables, rather than simply selecting winning candidates—an ability that is particularly relevant to modeling phonetic data, which always involves variation.

If maxent phonetic grammars are paired with constraints with continuous violation functions (such as the DURATION constraint family), the languages that these grammars are capable of generating are restricted in their typology, in the sense that the framework and constraint families make empirical predictions about how phonetics variables like duration can and cannot pattern. Some of these typological predictions are very general, and independent of the specific constraints used, while others depend on choices regarding what constraints families are used, and especially on how constraints assign violations.

7.2.1. The Consistent Variation Hypothesis

One very general prediction is dubbed the Consistent Variation Hypothesis (section 4.1.5.1): because the same constraints govern phonologically-conditioned variation and "random" free-variation, these two should be correlated. Classes of sounds or prosodic constituents which show a large degree of free variation in some phonetic variable should be just the classes of sounds for which that phonetic variable responds more readily to external conditioning factors. This prediction is wide in its scope, and is empirically testable.

Experimental results from this dissertation provide preliminary support for consistent variation in the duration domain (section 5.4.7): when paired natural classes which differ only in a single phonological feature or environment are compared, the class which shows more random durational variation across tokens is nearly always the category whose duration is more effected by various other phonological factors.

7.2.2. Constraint synergy

Another such prediction is that phonetic effects should combine *synergistically*: when a number of conditioning factors all induce similar phonetic effects—for example, are all associated with shortening—the more of these factors are already present, the smaller the effect of adding another such factor will be.[52] This is because candidates can satisfy multiple constraints simultaneously (just as they can in phonological constraint grammars), alleviating the need for each constraint to have an independent contribution to an effect size in the case where both are present.

This concept can be used to explain the Hyperadditive Lengthening Generalization: if all or most factors effecting duration are shortening effects, when two or more of these apply in tandem, synergistic shortening results in under-shortening of these shortest cases, which is equivalent to the hyperadditive lengthening interaction (section 5.5.1.2).

[52] The empirical testability of this feature of maxent grammars relies on knowing what is the base case for any given phonetic effect (for example, knowing whether speakers are lengthening phrase-final syllables, shortening phrase-medial ones, or both). Unfortunately, this is rarely clear *a priori*.

7.2.3. Phonetic constraints and phonetic distributions

While all maxent phonetic grammars output probability distributions over phonetic realizations, the shapes of these distributions depend on the choices of constraints, and especially the formulation of these constraints' violation functions. Constraints with parabolic violation functions centered on a target, like the DURATION constraints proposed by Flemming (2001) have the interesting property that maxent grammars using sets of these constraints always output normal distributions.

Because the more general constraint family STRETCH and SQUEEZE involves asymmetrical constraints, the distributions predicted by grammars of these constraints are not always normal, but can be asymmetrical. However, this asymmetry is itself rule-governed: relatively longer categories, since they are subject to more or more highly weighted STRETCH constraints and to fewer or less highly weighted SQUEEZE constraints, will have, if anything, relatively greater skewness (section 4.4.2) than comparable shorter categories.

This pattern is in fact seen in experimental results, in that the duration means of the samples in the experiment are significantly positively correlated with their kurtoses (section 5.4.6).

7.2.4. Phonetic learning

Maxent phonetic grammars are learnable, at least in the narrow sense that it is possible to algorithmically determine their parameters so as to best fit a set of a phonetic training data. For constraints which involve targets, these targets can be learned simultaneously with constraint weights.

Multiple local optima seem to exist in the learning space, and as a result the learner may not always learn the same grammar given the same data. However, for reasonable constraint sets, the

various local optima found are qualitatively similar, and produce similar fits to the data, suggesting that phonetic grammars simply have multiple ways of predicting a given set of data.

The targets which end up being learned for the constraints of grammar fragments for English front vowel duration are, at first glance, counterintuitive: unlike in prior work on targeted phonetic constraints, the constraint targets cannot be phonetic targets in the traditional sense, since they are in many cases quite extreme, or even negative numbers. This is partly because the grammar fragments are complex enough that many constraints are active in determining the duration of any particular input to the grammar, such that the output is always a function of multiple constraints, and not any single constraint. It is also partly because, unlike harmonic grammars, maxent grammars are responsible for modeling the variation seen in the training data, and in some cases they can only do so by using very small constraint weights coupled with extremely long or extremely short constraint targets.

7.3. New research directions

This dissertation, like many, raises far more questions than it answers. I here hint at how some of the research threads that have been picked up might be followed, and invite (and encourage) readers with an interest in theoretical phonetics to follow them.

7.3.1. Empirical

The empirical predictions made in this dissertation have, as yet, only weak support. The Hyperadditive Lengthening Generalization, for example, is based on several interactions between pairs of effects influencing duration, but the way most pairs of phonetic effects interact is still unknown. Filling these empirical gaps would help to either support or to reject this generalization.

More broadly, the interplay between multiple phonetic processes of all kinds is not yet well understood. When interactions between independent variables are found in the empirical phonetics literature, they are often merely statistical footnotes in experimental work whose goal is to establish the importance of each variable independently. For linguists interested in generative models of phonetics, however, an empirical understanding about the way multiple factors governing the same phonetic variable interact with each other, and especially any generalizations that can be made about such interactions, is of tantamount importance to understanding what kind of generative model of phonetics is needed.

Another prediction, the Consistent Variation Hypothesis—a broad prediction of maxent phonetic constraint grammars—is wide open for empirical testing, not only for phonetic duration, but for every other conceivable phonetic variable.

7.3.2. Theoretical

The learning of maxent phonetic constraint grammars is in its infancy, this dissertation serving only as a proof of concept, even when it comes to the duration domain. More realistic maxent grammars for duration would be more like the harmonic grammars used by Katz (2010), including constraints on the durations of both segments and on prosodic constituents larger than segments, such as syllables or feet, predicting the durations of multiple segments at once. Predicting multiple duration values would require a multi-dimensional candidate set, making GEN a bit more complex.

The best way to formulate the phonetic constraints themselves is a topic ripe for additional research: unlike phonological constraints, phonetic constraints are relatively new, and which constraints should be used and how they should assign violations to candidates is an open theoretical question. For example, while parabolic and hemiparabolic violation functions have

been used throughout this dissertation, many other functions are plausible (cf. Windmann et al., 2015), could easily be employed in the maxent phonetics framework, and once so employed would make their own empirically testable predictions about patterns in the distributions of phonetic variables (section 4.1.5). Investigation of this sort could also proceed in the opposite direction: if the distribution of some phonetic variable is already known to be interesting in some way (for example by being non-normal, or by showing more variation in one phonological category or in one context than in another) then these facts could be used as a basis for reverse engineering what sort of constraint or constraints would need to govern that phonetic variable in a phonetic grammar.

7.4. Outlook

While many of the particulars are as yet unclear, one thing at least seems certain: the maxent framework provides structural linguists with a powerful formalism for developing generative, restrictive, learnable grammars which govern not only the systematic relationships between levels of phonological representation, but also between these discrete phonological representations and quasi-continuous phonetic ones.

Appendix: experimental results

This appendix provides aggregate statistics for the tokens in each of the 128 experimental conditions (see Chapter 5).

Phrase-Medial, Unaccented

Vowel	Coda	Word	N	Duration (s)			Log Duration (s)		
				mean	stdev	skew	mean	stdev	skew
æ	_d	bad	7	0.187	0.043	-0.229	-1.703	0.246	-0.466
		mad	4	0.148	0.029	0.252	-1.933	0.195	-0.018
	_t	bat	7	0.156	0.040	-0.445	-1.900	0.298	-0.986
		mat	8	0.135	0.020	0.457	-2.015	0.146	0.252
	_ts	bats	10	0.165	0.019	0.301	-1.808	0.115	0.250
		mats	11	0.123	0.022	0.566	-2.107	0.171	0.195
ɛ	_d	bed	7	0.123	0.017	-0.473	-2.106	0.144	-0.859
	_t	bet	5	0.124	0.029	-0.121	-2.116	0.243	-0.205
		met	8	0.095	0.024	-0.348	-2.388	0.284	-0.881
	_ts	bets	8	0.128	0.029	0.354	-2.083	0.229	0.017
ɪ	_d	bid	7	0.106	0.040	-0.173	-2.332	0.425	-0.417
		mid	9	0.083	0.021	0.021	-2.519	0.265	-0.425
	_t	bit	12	0.107	0.030	0.648	-2.274	0.271	0.205
		mitt	6	0.090	0.019	0.425	-2.434	0.212	0.290
	_ts	bits	12	0.089	0.017	0.229	-2.437	0.189	-0.178
		mitts	5	0.054	0.011	0.214	-2.935	0.200	-0.094

Vowel	Coda	Word	N	Duration (s)			Log Duration (s)		
				mean	stdev	skew	mean	stdev	skew
eɪ	_∅	bay	4	0.125	0.005	0.000	-2.080	0.040	0.000
		may	11	0.113	0.027	0.689	-2.207	0.227	0.317
	_d	bade	4	0.183	0.063	1.092	-1.749	0.301	1.037
		made	11	0.126	0.032	1.195	-2.097	0.229	0.926
	_t	bait	9	0.154	0.026	0.321	-1.888	0.167	0.214
		mate	4	0.108	0.011	-0.652	-2.236	0.106	-0.749
	_ts	baits	10	0.140	0.020	0.356	-1.976	0.141	0.063
		mates	8	0.099	0.018	0.417	-2.334	0.184	0.278
i	_∅	bee	8	0.130	0.020	0.496	-2.053	0.151	0.136
		me	8	0.113	0.027	-0.297	-2.215	0.268	-1.097
	_d	bead	8	0.144	0.031	0.534	-1.961	0.211	0.252
		mead	7	0.157	0.029	-0.077	-1.868	0.192	-0.368
	_t	beat	11	0.125	0.025	-0.263	-2.105	0.213	-0.613
		meat	10	0.116	0.019	-0.158	-2.167	0.164	-0.273
	_ts	beats	10	0.118	0.015	-0.520	-2.146	0.136	-0.660
		meats	9	0.117	0.024	-0.255	-2.169	0.224	-0.855

Phrase-Medial, Accented

Vowel	Coda	Word	N	Duration (s)			Log Duration (s)		
				mean	stdev	skew	mean	stdev	skew
æ	_d	bad	11	0.232	0.033	0.909	-1.472	0.135	0.582
		mad	12	0.208	0.054	0.906	-1.600	0.243	0.632
	_t	bat	14	0.179	0.031	0.353	-1.733	0.169	0.130
		mat	12	0.150	0.023	-0.488	-1.909	0.164	-0.706
	_ts	bats	12	0.188	0.023	0.102	-1.682	0.125	-0.084
		mats	14	0.149	0.017	0.586	-1.908	0.114	0.360
ɛ	_d	bed	14	0.176	0.047	0.952	-1.768	0.251	0.473
	_t	bet	12	0.153	0.031	0.351	-1.901	0.204	-0.050
		met	12	0.100	0.020	-0.883	-2.328	0.242	-1.538
	_ts	bets	11	0.124	0.017	0.816	-2.092	0.129	0.556
ɪ	_d	bid	12	0.140	0.032	0.169	-1.991	0.230	-0.150
		mid	10	0.108	0.038	0.288	-2.292	0.371	-0.342
	_t	bit	8	0.114	0.025	-0.064	-2.198	0.225	-0.284
		mitt	8	0.084	0.027	1.690	-2.523	0.272	1.206
	_ts	bits	13	0.122	0.032	0.099	-2.145	0.280	-0.529
		mitts	6	0.085	0.022	1.116	-2.497	0.240	0.729

Vowel	Coda	Word	N	Duration (s)			Log Duration (s)		
				mean	stdev	skew	mean	stdev	skew
eɪ	_Ø	bay	12	0.210	0.050	0.198	-1.592	0.246	-0.266
		may	10	0.185	0.046	0.973	-1.716	0.232	0.589
	_d	bade	11	0.214	0.038	0.435	-1.558	0.176	0.273
		made	14	0.190	0.062	0.852	-1.709	0.306	0.505
	_t	bait	14	0.178	0.023	0.081	-1.735	0.131	-0.109
		mate	12	0.135	0.022	1.020	-2.014	0.158	0.548
	_ts	baits	12	0.162	0.028	0.715	-1.832	0.168	0.245
		mates	13	0.126	0.022	0.235	-2.087	0.174	-0.110
i	_Ø	bee	12	0.188	0.052	0.174	-1.714	0.291	-0.367
		me	10	0.189	0.050	0.126	-1.703	0.282	-0.493
	_d	bead	11	0.183	0.038	-0.151	-1.724	0.224	-0.606
		mead	11	0.184	0.046	0.439	-1.721	0.247	0.123
	_t	beat	11	0.148	0.019	-0.633	-1.916	0.137	-0.806
		meat	11	0.128	0.023	-0.096	-2.073	0.194	-0.684
	_ts	beats	12	0.154	0.027	0.272	-1.889	0.177	0.139
		meats	13	0.125	0.021	0.382	-2.096	0.170	0.014

Phrase-Final, Unaccented

Vowel	Coda	Word	N	Duration (s)			Log Duration (s)		
				mean	stdev	skew	mean	stdev	skew
æ	_d	bad	9	0.249	0.056	-1.015	-1.423	0.274	-1.603
		mad	10	0.216	0.057	1.588	-1.560	0.228	1.275
	_t	bat	9	0.196	0.044	0.190	-1.659	0.236	-0.358
		mat	9	0.162	0.052	-0.660	-1.892	0.432	-1.625
	_ts	bats	4	0.228	0.033	0.579	-1.491	0.143	0.441
		mats	5	0.110	0.036	-0.550	-2.277	0.400	-0.888
ɛ	_d	bed	11	0.173	0.042	1.749	-1.782	0.214	1.043
	_t	bet	12	0.162	0.035	0.258	-1.846	0.218	-0.213
		met	10	0.122	0.030	-0.635	-2.140	0.283	-1.162
	_ts	bets	7	0.170	0.033	-1.150	-1.795	0.227	-1.409
ɪ	_d	bid	9	0.160	0.048	-0.048	-1.883	0.327	-0.482
		mid	7	0.127	0.037	0.277	-2.110	0.304	-0.232
	_t	bit	7	0.149	0.020	0.178	-1.915	0.136	0.038
		mitt	3	0.113	0.033	-0.295	-2.226	0.320	-0.458
	_ts	bits	9	0.112	0.030	0.449	-2.223	0.268	0.142
		mitts	4	0.108	0.027	0.657	-2.260	0.238	0.474

Vowel	Coda	Word	N	Duration (s)			Log Duration (s)		
				mean	stdev	skew	mean	stdev	skew
eɪ	_Ø	bay	6	0.310	0.049	0.391	-1.183	0.154	0.282
		may	4	0.218	0.045	-0.426	-1.549	0.221	-0.619
	_d	bade	8	0.215	0.039	0.876	-1.552	0.174	0.442
		made	10	0.198	0.060	0.826	-1.663	0.292	0.206
	_t	bait	6	0.175	0.028	-0.383	-1.756	0.165	-0.552
		mate	7	0.156	0.041	-0.548	-1.903	0.306	-0.963
	_ts	baits	8	0.150	0.031	-1.201	-1.922	0.248	-1.610
		mates	8	0.138	0.023	-0.452	-1.997	0.176	-0.558
i	_Ø	bee	10	0.221	0.028	1.176	-1.516	0.119	0.806
		me	3	0.210	0.022	-0.595	-1.566	0.107	-0.616
	_d	bead	13	0.201	0.019	0.422	-1.610	0.093	0.220
		mead	7	0.184	0.033	0.747	-1.706	0.172	0.660
	_t	beat	9	0.161	0.024	-1.554	-1.839	0.176	-1.867
		meat	12	0.158	0.034	-0.559	-1.874	0.245	-1.320
	_ts	beats	6	0.130	0.026	-0.521	-2.065	0.227	-1.015
		meats	8	0.125	0.032	-0.268	-2.116	0.282	-0.641

Phrase-Final, Accented

Vowel	Coda	Word	N	Duration (s)			Log Duration (s)		
				mean	stdev	skew	mean	stdev	skew
æ	_d	bad	12	0.313	0.055	1.030	-1.174	0.165	0.798
		mad	14	0.244	0.037	0.316	-1.423	0.151	0.134
	_t	bat	13	0.209	0.036	0.391	-1.578	0.170	0.071
		mat	11	0.170	0.042	-0.664	-1.808	0.288	-1.249
	_ts	bats	10	0.221	0.032	0.347	-1.519	0.144	0.014
		mats	11	0.168	0.050	-0.039	-1.832	0.325	-0.624
ɛ	_d	bed	10	0.235	0.052	0.355	-1.472	0.219	0.145
	_t	bet	10	0.157	0.031	-0.728	-1.876	0.226	-1.147
		met	13	0.105	0.022	0.479	-2.278	0.210	0.041
	_ts	bets	8	0.132	0.024	-0.237	-2.039	0.189	-0.563
ɪ	_d	bid	12	0.193	0.040	0.856	-1.667	0.199	0.259
		mid	10	0.162	0.027	-0.194	-1.834	0.173	-0.601
	_t	bit	10	0.137	0.037	-0.132	-2.029	0.297	-0.727
		mitt	6	0.090	0.020	-0.125	-2.434	0.234	-0.394
	_ts	bits	8	0.153	0.044	0.060	-1.926	0.311	-0.503
		mitts	7	0.099	0.030	0.425	-2.363	0.304	0.069

Vowel	Coda	Word	N	Duration (s)			Log Duration (s)		
				mean	stdev	skew	mean	stdev	skew
eɪ	_Ø	bay	11	0.349	0.057	0.333	-1.065	0.161	0.112
		may	13	0.250	0.062	0.482	-1.417	0.248	-0.097
	_d	bade	11	0.291	0.040	0.025	-1.244	0.137	-0.132
		made	13	0.232	0.055	1.704	-1.486	0.208	1.222
	_t	bait	10	0.204	0.031	0.746	-1.601	0.148	0.549
		mate	13	0.151	0.045	-0.421	-1.952	0.376	-1.436
	_ts	baits	12	0.178	0.027	-0.503	-1.736	0.160	-0.873
		mates	10	0.141	0.021	0.271	-1.970	0.147	0.049
i	_Ø	bee	12	0.313	0.064	1.883	-1.181	0.180	1.397
		me	12	0.247	0.064	-0.057	-1.435	0.274	-0.384
	_d	bead	12	0.253	0.041	-0.177	-1.390	0.168	-0.352
		mead	9	0.237	0.045	0.276	-1.458	0.188	0.084
	_t	beat	7	0.163	0.031	0.019	-1.834	0.194	-0.143
		meat	8	0.146	0.028	0.497	-1.941	0.189	0.228
	_ts	beats	10	0.163	0.018	0.229	-1.821	0.113	-0.071
		meats	13	0.144	0.025	0.702	-1.952	0.168	0.355

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control, 19*(6), 716-723.

Anderson, M., Pierrehumbert, J., & Liberman, M. (1984). Synthesis by rule of English intonation patterns. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84*. (Vol. 9, pp. 77-80). IEEE.

Aylett, M., & 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., Turk, A. (2006) Language Redundancy Predicts Syllabic Duration and the Spectral Characteristics of Vocalic Syllable Nuclei, *Journal of the Acoustical Society of America 119*: 3048-58

Beckman, M. E., & Pierrehumbert, J. B. (1986). Japanese prosodic phrasing and intonation synthesis. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics* (pp. 173-180). Association for Computational Linguistics.

Best, Catherine T. (1995). A Direct Realist View of Cross-Language Speech Perception. *Speech perception and linguistic experience: Issues in cross-language research*, 171-204.

Baese-Berk, M., & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language & Cognitive Processes 24*: 527-554.

Boersma, Paul (1998) Functional Phonology: Formalizing the interactionsbetween articulatoryand perceptual drives. Doctoral dissertation, University of Amsterdam. The Hague: Holland Academic Graphics.

Boersma, P. (2003, January). Stochastic Optimality Theory. In *Meeting of the Linguistic Society of America*.

Boersma, P., & Pater, J. (2008). Convergence properties of a gradual learning algorithm for Harmonic Grammar.

Boersma, P. (2009). Cue constraints and their interactions in phonological perception and production. In *Phonology in perception, 15*, 55-110.

Boersma, P. (2011). A programme for bidirectional phonology and phonetics and their acquisition and evolution. *Bidirectional optimality theory, 180*, 33.

Braver, A. (2013). *Degrees of incompleteness in neutralization: Paradigm uniformity in a phonetics with weighted constraints*. Rutgers The State University of New Jersey-New Brunswick.

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

Byrd, D. (1996b). A phase window framework for articulatory timing. *Phonology, 13*(02), 139-169.

Byrd, D., & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics, 26*(2), 173-199.

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

Byrd, D., & Tan, C. C. (1996). Saying consonant clusters quickly. *Journal of Phonetics, 24*(2), 263-282.

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of phonetics, 27*(2), 207-229.

Choi, J., Kim, S., & Cho, T. (2016). Phonetic encoding of coda voicing contrast under different focus conditions in L1 vs. L2 English. *Frontiers in psychology, 7*.

Chomsky, N., & Halle, M. (1968). The sound pattern of English.

Cooper, W. E., & Danly, M. (1981). Segmental and temporal aspects of utterance-final lengthening. *Phonetica, 38*(1-3), 106-115.

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Current results. *The Journal of the Acoustical Society of America, 83*(4), 1553-1573.

Daland, R. (2015). Long words in maximum entropy phonotactic grammars. *Phonology, 32*(03), 353-383.

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.

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing features of random fields. *IEEE transactions on pattern analysis and machine intelligence, 19*(4), 380-393.

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

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. *The Handbook of Phonetic Sciences, Second Edition*, 521-602.

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

Flemming, E. (2004). Contrast and perceptual distinctiveness. *Phonetically-based phonology*, 232-276.

Flemming, E., & Cho, H. (2017). The phonetic specification of contour tones: evidence from the Mandarin rising tone. *Phonology, 34*(1), 1-40.

Fougeron, C. E. C., & Jun, S. A. (1998). Rate effects on French intonation: Prosodic organization and phonetic realization. *Journal of Phonetics, 26*(1), 45-69.

Fowler, C. A. (1988). Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech, 31*(4), 307-319.

Gafos, A. I. (2002). A grammar of gestural coordination. *Natural Language & Linguistic Theory, 20*(2), 269-337.

Gay, T., Ushijima, T., Hirose, H., & Cooper, F. S. (1974). Effect of Speaking Rate on Labial Consonant-Vowel Articulation. *The Journal of the Acoustical Society of America, 55*(2), 385-385.

Goldwater, Sharon and Mark Johnson (2003) Learning OT constraint rankings using a Maximum Entropy model. *Proceedings of the Workshop on Variation within Optimality Theory*, Stockholm University, 2003.

Halle, M. (1978). Knowledge unlearned and untaught: What speakers know about the sounds of their language. In M. Halle, J. Bresnan, and G. Miller (Eds.), *Linguistic Theory and Psychological Reality*, pp. 294–303. MIT Press.

Hawkins, S., & Warren, P. (1994) Phonetic influences on the intelligibility of conversational speech. *Journal of Phonetics 22*: 493-511.

Hayes, Bruce & Russell Schuh (2017). *Metrical structure and sung rhythm of the Hausa Rajaz*. Ms., Department of Linguistics, UCLA, Los Angeles, CA.

Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry 39*: 379-440.

Jakobson, R., & Waugh, L. R. (1979). *The sound shape of language*. Indiana University Press.

Katz, J. J. I. (2010). Compression effects, perceptual asymmetries, and the grammar of timing (Doctoral dissertation, Massachusetts Institute of Technology).

Katz, J. (2012). Compression effects in English. *Journal of Phonetics, 40*(3), 390-402.

Kawahara, S., & Braver, A. (2013). The Phonetics of Multiple Vowel Lengthening in Japanese. *Open Journal of Modern Linguistics, 3*(02), 141.

Keating, P. A. (1979). *A phonetic study of a voicing contrast in Polish*. Unpublished PhD dissertation, Brown University.

Keating, P. A. (1985). Universal phonetics and the organization of grammars. In *Phonetic linguistics: Essays in honor of Peter Ladefoged*, ed. by Victoria A. Fromkin, 115-32.

Keating, P. (1990a). The window model of coarticulation: articulatory evidence. *Papers in laboratory phonology I*, 26, 451-470.

Keating, P. (1990b). Phonetic representations in a generative grammar. *Journal of phonetics, 18*(3), 321-334.

Keating, P., Cho, T., Fougeron, C., & Hsu, C. S. (2003). Domain-initial articulatory strengthening in four languages. *Papers in laboratory phonology VI*, 145-163.

Klatt, D. H. (1973a). Durational characteristics of prestressed word-initial consonant clusters in English. MIT Research Laboratory of Electronics, *Quarterly Progress Report, 108*, 253-260.

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

Klatt, D. (1974). The duration of [s] in English words. *Journal of Speech, Language and Hearing Research, 17*(1), 51.

Klatt, D. H. (1975). Vowel Lengthening is Syntactically Determined in a Connected Discourse. *Journal of phonetics, 3*(3), 129-140.

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. (1982, May). The Klattalk text-to-speech conversion system. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.* (Vol. 7, pp. 1589-1592). IEEE.

Legendre, G., Miyata, Y., & Smolensky, P. (1990). *Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations*. University of Colorado, Boulder, Department of Computer Science.

Lehiste, I. (1970). *Suprasegmentals*. The MIT Press.

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

Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *The Journal of the Acoustical Society of America, 54,* 1228.

Lehiste, I. (1975a). Some factors affecting the duration of syllabic nuclei in English. In *Proceedings of the First Salzburg Conference on Linguistics* (pp. 81-104).

Lehiste, I. (1975b). The phonetic structure of paragraphs. In *Structure and process in speech perception* (pp. 195-206). Springer Berlin Heidelberg.

Li, A., & Post, B. (2014). L2 acquisition of prosodic properties of speech rhythm. *Studies in Second Language Acquisition, 36*(02), 223-255.

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

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

Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech, 47*(2), 109-138.

Luce, P. A., & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *The Journal of the Acoustical Society of America, 78*(6), 1949-1957.

Maddieson, I. (2004). Timing and alignment: A case study of Lai. *Language and Linguistics*, 5(4), 729-755.

Mayer, C., Gick, B., & Ferch, E. (2009). Talking while chewing: Speaker response to natural perturbation of speech. *Canadian Acoustics*, 37(3), 144-145.

McCarthy, J. J. (2000). Harmonic serialism and parallelism.

Moreton, E. (2004). Realization of the English postvocalic [voice] contrast in F 1 and F 2. *Journal of Phonetics, 32*(1), 1-33.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133-142

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America, 32*(6), 693-703.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* (Doctoral dissertation, Massachusetts Institute of Technology).

Pierrehumbert, J. (1981). Synthesizing intonation. *The Journal of the Acoustical Society of America*, *70*, 985.

Prince, Alan (1980) A metrical theory for Estonian quality. *Linguistic Inquiry, 11* (1980), pp. 511-562.

Remijsen, B., & Gilley, L. (2008). Why are three-level vowel length systems rare? Insights from Dinka (Luanyjang dialect). *Journal of Phonetics*, *36*(2), 318-344.

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.

Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). FAVE (forced alignment and vowel extraction) program suite. *URL http://fave.ling.upenn.edu*.

Scarborough, R.A. (2004) *Coarticulation & the structure of the lexicon*. Ph.D. dissertation, UCLA, Los Angeles, CA.

Small, A. M., & Campbell, R. A. (1962). Temporal differential sensitivity for auditory stimuli. *The American journal of psychology*, *75*(3), 401-410.

Smith, C. L. (1993). *The timing of vowel and consonant gestures.* (Doctoral dissertation, Yale University).

Stevens, S. S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, *53*(3), 329-353.

Van Summers, W. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *The Journal of the Acoustical Society of America*, *82*(3), 847-863.

Turk, A., & Sawusch, J. R. (1993). On the perceptual integrality of duration and amplitude cues to stress. *The Journal of the Acoustical Society of America*, *93*(4), 2371-2371.

Turk, A. E., & White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics*, *27*(2), 171-206.

Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, *35*(4), 445-472.

Umeda, N. (1975). Vowel duration in American English. *The Journal of the Acoustical Society of America*, *58*(2), 434-445.

Van Santen, J. P. (1992). Contextual effects on vowel duration. *Speech communication*, *11*(6), 513-546.

Van Santen, J. P., Shih, C., Möbius, B., Tzoukermann, E., & Tanenblatt, M. (1997). Multi-lingual duration modeling. In *EUROSPEECH*.

Vogt, F., Guenther, O., Hannam, A., van den Doel, K., Lloyd, J., Vilhan, L., ... & Derrick, D. (2005). ArtiSynth designing a modular 3D articulatory speech synthesizer. *The Journal of the Acoustical Society of America*, 117(4), 2542-2542.

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

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America, 91*, 1707.

Wright (2004) R. Wright, Factors of lexical competition in vowel articulation. *Papers in Laboratory Phonology 6*, Cambridge University Press.

Zhang, J. (2007). Constraint weighting and constraint domination: a formal comparison. *Phonology, 24*(3), 433.

Zsiga, E. C. (2000). Phonetic alignment constraints: Consonant overlap and palatalization in English and Russian. *Journal of Phonetics, 28*(1), 69-102.