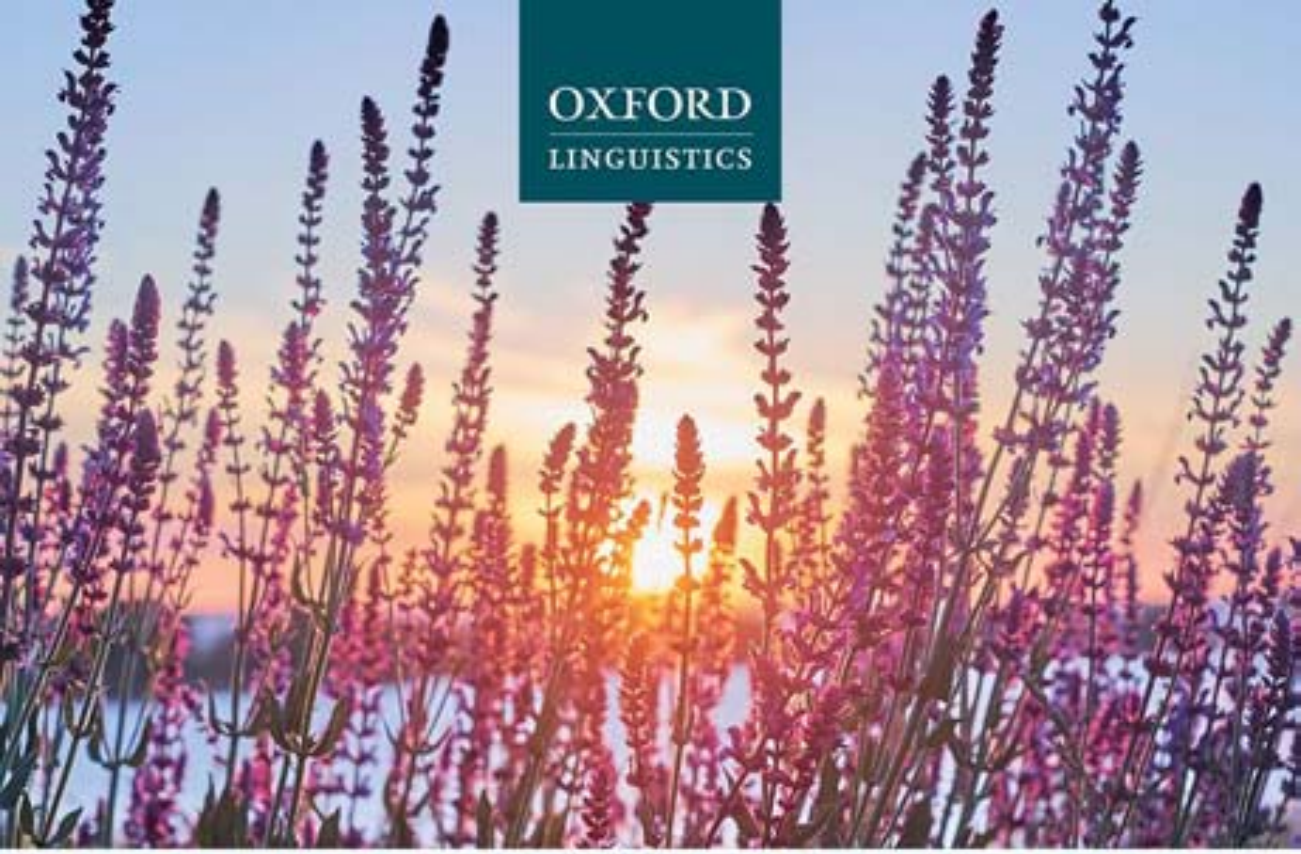

Speech Timing

Implications for Theories of Phonology, Phonetics, and Speech Motor Control

ALICE TURK
AND
STEFANIE SHATTUCK-HUFNAGEL

OXFORD STUDIES IN PHONOLOGY & PHONETICS

Speech Timing

OXFORD STUDIES IN PHONOLOGY AND PHONETICS

PUBLISHED

1
Morphological Length and Prosodically Defective Morphemes
Eva Zimmermann

2
The Phonetics and Phonology of Geminate Consonants
Edited by Haruo Kubozono

3
Prosodic Weight: Categories and Continua
Kevin M. Ryan

4
Phonological Templates in Development
Marilyn May Vihman

5
Speech Timing: Implications for Theories of Phonology, Phonetics, and Speech Motor Control
Alice Turk and Stefanie Shattuck-Hufnagel

IN PREPARATION

Phonological Specification and Interface Interpretation
Edited by Bert Botma and Marc van Oostendorp

Doing Computational Phonology
Edited by Jeffrey Heinz

Intonation in Spoken Arabic Dialects
Sam Hellmuth

Synchronic and Diachronic Approaches to Tonal Accent
Edited by Pavel Iosad and Björn Köhnlein

The Structure of Nasal-Stop Inventories
Eduardo Piñeros

Speech Timing

Implications for Theories of Phonology, Phonetics, and Speech Motor Control

ALICE TURK
AND
STEFANIE SHATTUCK-HUFNAGEL

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2020

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2019945699

ISBN 978–0–19–879542–1

Printed and bound in Great Britain by
Clays Ltd, Elcograf S.p.A.

Contents

Series preface ix
Acknowledgments xi
List of figures and tables xiii
List of abbreviations xv

1. Introduction 1

2. Articulatory Phonology/Task Dynamics 8
   2.1 Introduction 8
   2.2 The dual function of gestures within Articulatory Phonology: contrast and constriction formation 10
   2.3 Using mass–spring systems to model gestural movement in TD 11
   2.4 Gestural control of individual articulators, and gestural activation 14
   2.5 Timing Control in AP/TD 20
   2.6 Key features of AP/TD 38
   2.7 Advantages of the AP/TD framework 43
   2.8 Conclusion 47

3. Evidence motivating consideration of an alternative approach 49
   3.1 AP/TD default specifications require extensive modifications 50
   3.2 Relationships among distance, accuracy, and duration are not fully explained in AP/TD 53
   3.3 Distinct synchronous tasks cause spatial interference 57
   3.4 Issues not currently dealt with 62
   3.5 Summary 62

4. Phonology-extrinsic timing: Support for an alternative approach I 64
   4.1 Introduction 64
   4.2 A challenge to the use of mass–spring oscillators in the implementation of timing effects 67
   4.3 Evidence for the mental representation of surface durations 75
   4.4 Further evidence for general-purpose timekeeping mechanisms to specify durations and track time 90
   4.5 Conclusion 100

5. Coordination: Support for an alternative approach II 102
   5.1 Introduction 102
   5.2 Evidence consistent with AP/TD inter-planning-oscillator coupling, and alternative explanations 105

5.3 Evidence that requires the consideration of non-oscillatory approaches 112
5.4 Evidence that timing relationships in movement coordination are not always based on movement onsets 119
5.5 Possible mechanisms for endpoint-based timing and coordination 127
5.6 Planning inter-movement coordination and movement-onset timing 129
5.7 Summary of findings relating to movement coordination 130

6. The prosodic governance of surface phonetic variation: Support for an alternative approach III 132
6.1 Introduction 132
6.2 Evidence relating to Pi/Mu_T mechanisms for boundary- and prominence-related lengthening 133
6.3 Evidence relating to the coupled oscillator hierarchy mechanism for poly-subconstituent shortening 135
6.4 Evidence which challenges the use of oscillators in controlling overall speech rate 143
6.5 Summary 144

7. Evidence for an alternative approach to speech production, with three model components 146
7.1 Existing three-component models and some gaps they leave 150
7.2 Why the timing evidence presented earlier motivates the three components of the XT/3C approach, despite the gaps 158
7.3 Evidence for the separation between the Phonological and Phonetic Planning Components: Abstract symbols in Phonological Planning 162
7.4 The translation issue 171
7.5 Motivating the separation between Phonetic Planning and Motor-Sensory Implementation 178
7.6 Key components of the proposed model sketch 188

8. Optimization 190
8.1 General overview 194
8.2 Key features 195
8.3 What are the costs of movement? 199
8.4 Predictions of Stochastic Optimal Feedback Control Theory 214
8.5 Challenges for Optimal Control Theory approaches 218
8.6 Optimization principles in theories of phonology and phonetics 220
8.7 Conclusion 236

9. How do timing mechanisms work? 238
9.1 General-purpose timekeeping mechanisms 239

9.2 Lee's General Tau theory 256
9.3 Summary 262

10. A sketch of a Phonology-Extrinsic-Timing-Based Three-Component model of speech production 264
10.1 Phonological Planning 268
10.2 Phonetic Planning 298
10.3 Motor-Sensory Implementation 310
10.4 Summary and discussion 312

11. Summary and conclusion 313

References 321
Index 363

Series preface

Oxford Studies in Phonology and Phonetics provides a platform for original research on sound structure in natural language within contemporary phonological theory and related areas of inquiry such as phonetic theory, morphological theory, the architecture of the grammar, and cognitive science. Contributors are encouraged to present their work in the context of contemporary theoretical issues in a manner accessible to a range of people, including phonologists, phoneticians, morphologists, psycholinguists, and cognitive scientists. Manuscripts should include a wealth of empirical examples, where relevant, and make full use of the possibilities for digital media that can be leveraged on a companion website with access to materials such as sound files, videos, extended databases, and software.

This is a companion series to *Oxford Surveys in Phonology and Phonetics*, which provides critical overviews of the major approaches to research topics of current interest, a discussion of their relative value, and an assessment of what degree of consensus exists about any one of them. The *Studies* series will equally seek to combine empirical phenomena with theoretical frameworks, but its authors will propose an original line of argumentation, often as the inception or culmination of an ongoing original research program.

Based on a theory involving planning the time between acoustic landmarks, this book provides a model of speech production utilizing symbolic (non-gestural, without specific spatiotemporal content) phonological representations and phonology-extrinsic, non-speech-specific, general-purpose timing mechanisms that have sufficient flexibility to account for empirically documented timing behavior. The model takes into account a variety of sources of evidence, including listener-related factors, and presents an elegant model of speech production that involves separate planning components for phonology and phonetics, an Optimal Control Theory approach, and movement coordination based on movement endpoints and continuous tau coupling, rather than on movement onsets. This volume is a ground-breaking culmination of many years of research by the authors, and offers up much serious discussion for consideration, alongside pronounced challenges to competing theories of speech timing and task dynamics.

Andrew Nevins
Keren Rice

Acknowledgments

It is a deeply felt pleasure to acknowledge the many people who have helped to shape our thinking over the years of writing this book. We especially thank Elliot Saltzman and Dave Lee for their indefatigable patience in explaining the fine details of AP/TD and General Tau theory, sometimes multiple times. Louis Goldstein took time to answer our questions, and Ioana Chitoran, Fred Cummins, Jelena Krivokapić, Bob Ladd, and Juraj Šimko have read through parts of the book in their intermediate stages, and provided remarkably useful feedback. The comments of three anonymous reviewers and of Khalil Iskarous have also improved the book; one reviewer in particular provided extensive pages of extremely valuable comments (we think we know who you are, and we thank you!). Katrina Harris' work on the bibliography and Ada Ren-Mitchell's work on the index saved us many weeks of time; Katherine Demuth provided useful feedback. We thank Ken Stevens, who first sensitized us to the role of individual acoustic cues to distinctive features in speech processing, and who established an atmosphere of inquiry in the MIT Speech Communication Group that fostered critical and creative thinking, and inspired this collaborative effort. We are grateful to the Arts and Humanities Research Council, who funded the initial stages of research for the book (grant number AH/1002758/1 to the first author), and to the US National Science Foundation (grant numbers BCS 1023596, 1651190 and 1827598 to the second author).

Our friends and families also played a critical role in the completion of this volume, not least by asking us at regular intervals "Is Chapter 7 done yet?" We will not soon forget their support of the project, and their understanding of research visits, and an inordinate number of strategically timed Skype calls (made necessary by our collaboration across 3000 miles of Atlantic Ocean).

Finally, we would like to acknowledge the intellectual generosity of the developers of the AP/TD approach, whose work has been a model to us of how science should be done. All errors and omissions are, of course, our responsibility alone.

List of figures and tables

List of figures

2.1 Gestural scores for the words *mad* and *ban*, illustrating gestural activation intervals and their relative timing 15

2.2 Time functions of vocal tract variables, as measured using X-ray microbeam data, for the phrase *pea pots*, showing the in-phase (synchronous within twenty-five ms) coordination of the lip gesture for the /p/ in *pots* and the /a/ gesture for the vowel in *pots* 17

2.3 The coupling graph for *spot* (top) in which the tongue-tip (fricative) gesture and the lip-closure gesture are coupled (in-phase) to the tongue-body (vowel) gesture, while they are also coupled to each other in the antiphase mode 26

2.4 Coupling graphs for syllable sequences 27

2.5 Steady-state patterns of (slow) foot and (fast) syllable oscillators, with asymmetrical (foot-dominant) coupling between foot and syllable oscillators 31

2.6 A schematic gestural score for two gestures spanning a phrasal boundary instantiated via a π-gesture 32

3.1 Schematic diagrams of the templates for the four experimental conditions in Franz et al. (2001) 59

4.1 Start and end times (in milliseconds) of keypress movements for two repetitions of the same . . . *an epic* . . . sequence 69

4.2 Schematic illustration of data extraction 71

4.3 Scatter plots of protrusion duration interval versus consonant duration (left column); onset interval versus consonant duration (middle column), and offset interval versus consonant duration (right column) for lip-protrusion movements from four participants' /i_u/ sequences (shown in each of four rows) 72

4.4 Mean test vowel durations (in ms) in the baseline and three experimental conditions 82

4.5 Means and standard deviations for vowel duration as a function of lexical/morphological quantity—short stem in short grade (SS–SG), short stem in long grade (SS–LG), long stem in short grade (LS–SG), long stem in long grade (LS–LG)—and sentence context—(Medial, Final) 83

5.1 The distribution of keypress start and end times measured relative to the previous keypress 124
7.1 The utterance excerpt . . . *caught her* . . . produced by a Scottish female speaker from the Doubletalk corpus (Scobbie et al. 2013) 168
7.2 The utterance excerpt . . . *caught her* . . . produced by the same Scottish female speaker that produced . . . *caught her* . . . shown in Figure 7.1 169
7.3 A schematic diagram of the proposed XT/3C-v1 model 188
8.1 Prosodic structure as the interface between language and speech, illustrating some of the factors that influence Phonetic Planning 191
9.1 TauG guidance of the tongue when saying 'dad' 260
10.1 An example prosodic structure for *Mary's cousin George baked the cake* 271
10.2 A grid-like representation of prominence structure for one possible prosodification of *Mary's cousin George* 273
10.3 The complementary relationship between predictability (language redundancy) and acoustic salience yields smooth-signal redundancy (equal recognition likelihood throughout an utterance) 276
10.4 Factors that shape surface phonetics and their relationship to predictability, acoustic salience, and recognition likelihood 277
10.5 The utterance excerpt . . . *elf in the mirror* . . . spoken by a Southern British English speaker from the Doubletalk corpus (Scobbie et al. 2013) 309

List of tables

2.1 AP/TD tract variables and the model articulator variables that they govern 14
9.1 TauG guidance of the jaw, lips, and tongue in monologue recordings from the ESPF Doubletalk corpus 261

List of abbreviations

3C	three-component
AP/TD	Articulatory Phonology/Task Dynamics
C	consonant
DIVA	Directions into Velocities of Articulators
GLO	glottal aperture
GO	Gradient Order
GoDIVA	Gradient Order DIVA
LA	lip aperture
LP	lip protrusion
LTH	lower tooth height
MRI	Magnetic resonance imaging
OCT	Optimal Control Theory
OFCT	Optimal Feedback Control Theory
OT	Optimality Theory
SOFCT	Stochastic Optimal Feedback Control Theory
TADA	Task Dynamic Application
TD	Task Dynamics
TDCD	tongue-dorsum constriction degree
TDCL	tongue-dorsum constriction location
TTCD	tongue-tip constriction degree
TTCL	tongue-tip constriction location
TTCO	tongue-tip constriction orientation
V	vowel
VEL	velic aperture
VITE	Vector Integration To Endpoint
XT/3C	phonology-extrinsic-timing-based three-component approach

1
Introduction

This is a book about speech timing, and about the implications of speech timing patterns for the architecture of the speech production planning system. It uses evidence from motor timing variation to address the question of how words come to have such different acoustic shapes in different contexts. The book came about for two main reasons: First, it was written in reaction to a debate in the literature about the nature of phonological representations, which, together with a set of mechanisms that operate in relation to these representations, account for the range of systematic surface variation observed for phonologically equivalent forms. Phonological representations are proposed to be spatiotemporal by some (Fowler, Rubin, Remez and Turvey 1980), and symbolic (atemporal) by others (Henke 1966; Keating 1990; Fujimura 1992 et seq.; Guenther 1995 et seq.; Levelt et al. 1999; inter alia). The model of speech articulation which currently provides the most comprehensive account of systematic phonetic patterns, including timing, is a spatiotemporal approach called Articulatory Phonology (Browman and Goldstein 1985, 1992a; Saltzman, Nam, Krivokapić and Goldstein 2008). This model has many strengths, among them that it accurately captures a wide variety of complex characteristics of speech articulation (including smooth, single-peaked movement velocity profiles, coarticulation, and systematic effects of prosodic structure). However, its choice of spatiotemporal phonological representations and phonology-intrinsic timing mechanisms makes it structurally very different from approaches based on symbolic representations. Thus resolving the debate about spatiotemporal vs. symbolic representations has implications not only for phonological theory, but also for the architecture of the speech motor control system.

The second motivation was that our shared interest in timing patterns in speech led us to wonder about the type of motor control system that can best explain what is known about speech timing. Because one of the primary differences between symbolic and spatiotemporal approaches is in how they deal with timing, an evaluation of these theories in terms of available timing evidence simultaneously leads to answers to both questions, namely, about the nature of phonological representations, and about the type of motor

Speech Timing. First edition. Alice Turk and Stefanie Shattuck-Hufnagel

control system that can account for speech timing behavior. Evaluation of the Articulatory Phonology model from this point of view is presented in the first part of the book, and is made possible in large part by the exemplary explicitness of the Articulatory Phonology model. Because the lines of evidence presented here do not accord with Articulatory Phonology's spatiotemporal approach, in the second half the book we provide a sketch of a model of speech production based on symbolic phonological representations and phonology-extrinsic timing mechanisms that has the flexibility to account for known timing behavior.

As we have noted, the choice between symbolic, atemporal phonological representations and spatiotemporal phonological representations has several fundamental implications for the architecture of the speech production planning system. One of the most significant of these implications is the number of planning components that are required. In systems that include a planning component with symbolic (i.e. discrete, without specific spatiotemporal content) phonological representations (Henke 1966; Klatt 1976; Keating 1990; Shattuck-Hufnagel 1992; Shattuck-Hufnagel, Demuth, Hanson and Stevens 2011; Munhall 1993, Kingston and Diehl 1994; van Santen 1994; Guenther 1995; Clements and Hertz 1996; Levelt, Roelofs, and Meyer 1999; Fujimura 1992 et seq.; Goldrick, Baker, Murphy and Baese-Berk 2011; Houde and Nagarajan 2011; Perkell 2012; Lefkowitz 2017), a separate phonetic planning component is required to plan the details of surface timing and spatial characteristics for each context. These aspects of an utterance are not specified by the symbolic phonological representation. As a result, a separate phonetic planning process is required to map, or 'translate' from the representational vocabulary of abstract phonological symbols to a different (i.e. fully quantitative) representational vocabulary that can specify the physical form of an utterance. In contrast, the Articulatory Phonology system has a very different architecture; because its phonological representations are already spatiotemporal and fully quantitative in nature, it does not require a separate phonetic planning component. That is, because phonology in the Articulatory Phonology framework is already spatiotemporal, a single type of representational vocabulary is used throughout production, and this avoids the need for a separate phonetic component to plan the spatiotemporal details that are required to implement a symbolic phonological plan in a spoken utterance.

In addition to its implications for the architecture of the planning model, the choice between spatiotemporal and symbolic phonological representations also has important consequences for how these two approaches deal with timing issues. This is because time is *intrinsic* to phonological representations

in Articulatory Phonology, but is not part of phonology in symbol-based models. Because timing is intrinsic to the phonology, surface timing characteristics simply emerge from the phonological system, and are not explicitly represented, specified, or tracked. In contrast, in symbol-based approaches, timing is *extrinsic* to the phonological representations, so that surface timing characteristics must be explicitly planned in a separate phonetic planning component. These fundamental differences in how the two contrasting approaches deal with timing suggest an important criterion for comparing and evaluating them: that is, how well do they account for what is currently known about motor timing in general, and speech motor timing in particular? A large part of this book is devoted to such an evaluation, made possible by the explicit predictions of the Articulatory Phonology approach developed in the framework of Task Dynamics (Saltzman and Munhall 1989; Saltzman, Nam, Krivokapić and Goldstein 2008). This book presents a number of lines of evidence that are inconsistent with the Articulatory Phonology approach in particular and the phonology-intrinsic timing approach in general, and therefore suggest the need to consider a different approach based on phonology-extrinsic timing.

To begin, the first few chapters of the book lay out the key features of phonology-intrinsic-timing-based Articulatory Phonology in the Task Dynamics framework, and examine the oscillator-based mechanisms it uses. This model of speech production planning has evolved significantly over the years, under the influence of a number of important theoretical developments and observational findings. For example, the development of modern prosodic theory, with its hierarchy of prosodic constituents and prominence levels governing systematic patterns of duration variation (such as boundary-related lengthening, prominence-related lengthening, and poly-constituent shortening, see Chapters 6 and 10) led to the incorporation of planning oscillators for the syllable, the foot, and the phrase, and to the postulation of other timing-adjustment mechanisms for boundary- and prominence-related lengthening, and speech rate (Byrd and Saltzman 2003; Saltzman et al. 2008). These developments have resulted in a system which is significantly more complex than the original proposal, but provides a much-needed account of contextual variability, via the manipulation of the activation intervals for the spatiotemporal representations in different contexts. The added mechanisms begin to chip away at the initial attractive simplicity of its model, but don't undermine its core principles significantly.

The second and more telling part of the evaluation involves an examination of current evidence in the non-speech motor timing literature and in the

related speech literature that is currently not modeled within Articulatory Phonology. This evaluation reveals phenomena which appear to be incompatible with the phonology-intrinsic timing approach, and which therefore motivate the consideration of an alternative approach based on phonology-extrinsic timing. Some of these phenomena appear to require the representation and planning of surface timing characteristics. These are not consistent with phonology-intrinsic timing theories, because in such theories surface timing characteristics are not explicitly represented or planned as goals. Instead, they emerge from interacting components within the spatiotemporal phonological system. As a result, there is no mechanism for explicitly specifying surface timing, yet a number of observations suggest speakers can do this. Other phenomena suggest the involvement of general-purpose timekeeping mechanisms, which are not invoked in Articulatory Phonology because they are at odds with its phonology-intrinsic timing approach, in which the timing mechanism is specific to speech, and surface timing characteristics are emergent. Still other phenomena, relating to timing precision at movement endpoints, are also at odds with spatiotemporal phonological representations and thus have no principled explanation within Articulatory Phonology. In contrast, they can be straightforwardly explained in a three-component speech production system which combines symbolic phonological representations with separate phonetic planning and motor-sensory implementation components. The evaluation is then extended to the ways in which Articulatory Phonology has chosen to account for movement coordination and effects of prosodic structure on articulation, within the phonology-intrinsic timing framework. This evaluation in light of additional findings in the motor-control literature similarly suggests the need to consider approaches to coordination and suprasegmental structure that are different from the oscillator-based approach of Articulatory Phonology.

The results of this multipart evaluation in the first half of the book highlight the need to develop an alternative type of speech motor control model that can deal more straightforwardly with available motor timing evidence. Drawing on and extending existing proposals, the second half of the book sketches out a three-part model that includes a Phonological Planning Component, a separate Phonetic Planning Component, and a Motor–Sensory Implementation Component. This model has two goals: to provide a more complete description of the phonological planning process than is available in existing three-part symbol-based systems, and to provide an account of certain aspects of systematic variation in surface phonetic timing behavior that is not available either in existing three-component models or in Articulatory Phonology.

Like any model of speech production, including Articulatory Phonology, this alternative approach must meet a certain set of generally agreed-upon criteria for an adequate model. That is, it must make contact with the phonological information that specifies the lexical form of each word in the planned utterance; it must include some specification of the utterance-specific prosodic structure, including relative word prominence and the grouping of words into larger constituents; it must provide an account of the ways in which words and their sounds vary systematically in different contexts; and it must provide instructions to the articulatory control mechanisms that are adequate to match the observed quantitative facts about spoken utterances, such as appropriate articulator trajectories, acoustic patterns, and surface phonetic timing in both the acoustic and the articulatory domains.

The model proposed here builds on insights gained from existing symbol-based three-component models, but extends this approach to account more comprehensively for the details of surface phonetic variation, using general-purpose timing mechanisms that are extrinsic to the phonology. This extended three-component approach based on phonology-extrinsic general-purpose timing mechanisms follows traditional phonological theory in assuming symbolic phonological representations. However, early models based on symbolic representations, derived from Generative Phonology (Chomsky and Halle 1968), did not attempt to deal with the physical manifestation of speech—in a sense they stopped at the point when the surface form of an utterance was still symbolically represented. Later models in the Generative Phonology framework generate articulatory movements, but they do not provide a full account of surface timing. The approach advocated here develops these ideas further, by proposing a more comprehensive account of surface phonetic variability, including timing. In doing so, it incorporates some of the ideas in the existing literature (e.g. Keating 1990; Guenther 1995; Guenther, Ghosh, and Tourville 2006; Guenther 2016; Fujimura 1992, 2000 et seq.) but differs in three main ways that provide the flexibility necessary to account for the full range of systematic context-governed phonetic variability. First, the proposed approach provides an account of the types of task requirements that are specified in the Phonological Planning Component. This aspect of the proposed model is based on evidence highlighting the large number of contextual factors that can influence utterance-specific surface phonetic form, including timing characteristics, and must therefore be included in the phonological plan in non-quantitative (but sometimes relational) symbolic terms, for later development in quantitative terms to form the phonetic plan. In this proposed model, task requirements include the production of phonological contrasts

appropriately in different contexts (where context is defined by factors such as location in a hierarchical prosodic structure, relative speaking rate, dialect and idiolect, speaking style/situation, and others, see Turk and Shattuck-Hufnagel 2014), as well as the choice of appropriate symbolically expressed acoustic cues to meet these task requirements. The second difference from earlier approaches is that, although the proposal shares with other three-component models the assumption of phonology-extrinsic, general-purpose timekeeping mechanisms, it differs in its account of timing control. The proposed account is based on planning the timing between acoustic landmarks (Stevens 2002), and incorporates Lee's (1998) General Tau theory to plan appropriate movement velocity profiles and target-based movement coordination for the movements that achieve the landmarks. The computation of parameter values to be specified in the Phonetic Planning Component (including parameter values for timing) occurs via mechanisms proposed in Optimal Control Theory, to determine the optimum way of meeting utterance-specific goals specified in the Phonological Planning Component, at minimum cost. Optimal Control Theory models the choice of movements to satisfy multiple goals while economizing on effort, time, and other costs (cf. Nelson 1983; Lindblom 1990), and has been used in several recent models of surface timing variation in speech, e.g. Flemming (2001); Šimko and Cummins (2010, 2011); Katz (2010); Braver (2013); Lefkowitz (2017). Finally, like other three-component approaches, the model incorporates a Motor–Sensory Implementation Component to carry out the optimized instructions, consistent with the evidence that speakers track and adjust their movements when possible, to ensure that their acoustic goals are achieved. Such a component is widely agreed to be necessary, and specific proposals for how this component works have been advanced by Houde and Nagarajan (2011) and Guenther (2016); see also Hickok (2014).

The chapters that follow first lay out the major tenets of the Articulatory Phonology approach, and some of its remarkable successes in providing an account of speech phenomena such as coarticulation (Chapter 2). Because a full description of the theory is necessary in order to evaluate it in light of accumulating evidence about the nature of movement planning and motor control, this chapter provides a comprehensive description of its current state, with elements pulled together from disparate parts of the extensive relevant Articulatory Phonology literature. Several chapters are then devoted to explicating why, in our view, currently available evidence from motor timing in general and speech timing in particular suggests that an alternative model is needed (Chapters 3–6). Chapter 7 summarizes this timing evidence and

presents additional spatial evidence which suggests the value of developing a three-component model with phonology-extrinsic timing and abstract symbolic phonological representations. Chapters 8–9 present a number of components from the existing literature that could provide some of the pieces of such an alternative model. These components include Stochastic Optimal Feedback Control Theory (Todorov and Jordan 2002; Houde and Nagarajan 2011), and General Tau theory (Lee 1998). Chapter 10 draws all these elements together, providing a sketch of a phonology-extrinsic-timing-based three-component model of speech production planning, and Chapter 11 provides a summary of the main points made in the book.

Although many of the components of this alternative approach are drawn from the existing literature, they have not previously been combined into a model of acoustic and articulatory speech planning based on symbolic phonological representations and phonology-extrinsic timing, which can account for systematic surface phonetic variation in speech, including systematic surface timing patterns. That is one of the tasks that we have set ourselves in this book. The proposed model is still at the beginning stages of development, but we believe that its eventual computational implementation will provide a more principled and comprehensive account of phonetic behavior, and a more realistic account of speech production processing in general, than is currently available. We hope that other researchers will be inspired to consider whether their phonetic observations could be accounted for by such a model, and we look forward to some lively interactions.

2
Articulatory Phonology/Task Dynamics

2.1 Introduction

Articulatory Phonology, developed in the Task Dynamics framework (hereafter AP/TD), provided the first comprehensive model of phonology, speech articulation, and the connection between them (Browman and Goldstein 1985, 1989, 1990, 1992, 1995; Saltzman and Munhall 1989 and more recent developments) This theory, like any theory of speech motor control, faces the challenge of explaining the how 'the same sound' can be produced in systematically different ways in different contexts. AP/TD is based on the idea first c[illegible]n motor control, that "it is tendencies in dynamics—[illegible]rces and mutual influences among components tenc[illegible] or steady states—that are primarily responsible for [illegible]rocesses" (Kugler, Kelso, and Turvey 1980, p. 6), and [illegible]s therefore require minimal involvement of "intelligen[illegible]l planning and computation. While acknowledging the[illegible]c goals (tasks) in speech, the AP/TD approach attempts to reduce the burden of planning and regulation by adopting Fowler's (1977, 1980) proposal that phonological representations are (spatio)temporal, and thus that the timing of speech movements is intrinsic to the phonology. Because the dynamical spatiotemporal phonological representations determine the movements which shape the acoustic speech signal, there is no requirement for a separate phonetic planning component. And because the phonological representations are not symbolic, there is no requirement to translate from non-spatiotemporal, discrete symbols to quantitative, continuous articulatory movements. Thus the AP/TD approach addressed one of the most vexing problems in speech: The gap between symbolic representations in the mind and quantitative values in the speech signal.

Since the 1980s, the AP/TD approach has been further developed to account for many effects of context on speech articulation, including coarticulatory effects of adjacent context, and effects of prosodic position. Because of this, AP/TD currently provides the most comprehensive account of systematic spatiotemporal variability in speech. In doing so, it represents the standard

Speech Timing. First edition. Alice Turk and Stefanie Shattuck-Hufnagel

which any alternative theory of speech production must match or surpass, and provides a clear advantage over traditional phonological theories as a model of the speech production process. This chapter reviews the assumptions, mechanisms, and implications of this theory. Because one of its central tenets is that time is intrinsic to phonological representations, and a major goal of the book is to evaluate the theory in terms of its ability to account for timing behavior, a particular focus of the chapter is on the consequences of the commitment to phonology-intrinsic timing for the way speech timing phenomena are modeled. Understanding these aspects of the Articulatory Phonology approach is critical for identifying phenomena for which the model currently has no account (Chapters 3–6), and which have motivated the proposal of an alternative, phonology-extrinsic timing approach (laid out in Chapters 7–10).

Because there have been considerable developments to the AP/TD theory over the years in response to new data, the aim of this chapter is to pull together a full description of the structures and mechanisms which are proposed within the current theory to account for systematic variability in speech production. It is primarily based on a series of papers which together describe the current state of the theory: Browman and Goldstein (1985); Browman and Goldstein (1990a, b); Browman and Goldstein (1992a); Browman and Goldstein (2000); Browman and Goldstein (unpublished ms); Byrd and Saltzman (1998); Byrd and Saltzman (2003); Goldstein, Nam, Saltzman and Chitoran (2009); Krivokapić (2020); Nam, Goldstein, and Saltzman (2010); Nam, Saltzman, Krivokapić and Goldstein (2008); Saltzman and Byrd (2000); Saltzman, Löfqvist, and Mitra (2000); Saltzman and Munhall (1989); Saltzman, Nam, Goldstein, and Byrd (2006); and Saltzman, Nam, Krivokapić and Goldstein (2008). A computational implementation of the model has been developed by Nam, Browman, Goldstein, Proctor, Rubin, and Saltzman, and is described here: www.haskins.yale.edu/tada_download/index.php. Where appropriate, reference is also made to newer developments, e.g. Sorensen and Gafos (2016), which have not yet been incorporated in a fully working system.

This chapter first introduces gestures as units of contrast and constriction formation in AP/TD (Section 2.2), presents how a mass–spring system is used to model constriction formation (Section 2.3), and describes the function of gestures in controlling individual articulators, of gestural activation in controlling gestural movement (Section 2.4), and mechanisms for timing control (Section 2.5) within this system. It then summarizes the key features of the model (Section 2.6) and its advantages (Section 2.7). Finally, the last section looks ahead to the evidence laid out in Chapters 3–6, which motivates the alternative approach presented in the remainder of the book.

2.2 The dual function of gestures within Articulatory Phonology: contrast and constriction formation

In the AP/TD framework, basic units of phonological contrast and speech production are units of vocal tract constriction formation called gestures, e.g. tongue-tip closure, tongue-body opening. In this framework, the term 'gesture' has a very specific meaning, which is somewhat different from the common use of the term. Each gesture specifies a set of articulators responsible for achieving a particular constriction in the vocal tract. For example, the upper lip, lower lip, and jaw act together to form a bilabial constriction, and the tongue body, tongue tip, and jaw act together to form a tongue-tip constriction at the alveolar ridge. A central tenet of AP/TD is that gestures, in the technical AP/TD sense, have a dual function. On the one hand, they are contrastive phonological units, that is, units that distinguish word meanings. At the same time, they each specify a family of movement trajectories with the same constriction target, and describe how these trajectories unfold over time. In different utterances, the articulatory trajectories for a given gesture can be different for several reasons. The first is because a given gestural constriction represents the activity of a task-specific coordinative structure (cf. Kugler et al. 1980), and can thus be achieved through different contributions of individual articulators which make up the coordinative structure and act in a coordinated fashion to achieve the gestural goal. An example is when the upper lip, lower lip, and jaw act together to achieve a bilabial constriction gesture, and can compensate for one another when one of these articulators is perturbed from its normal pattern of activity (Folkins and Abbs 1975), or when one of its articulators is involved in the production of a different, overlapping gesture. Reasons for variability in the articulatory trajectories for a gesture include 1) differences in gestural starting position, e.g. because of a different preceding gesture produced with the same articulators; 2) differences in overlapping gestures; or 3) differences in how long a gesture is active, due either to differences in prosodic position, or to differences in speech rate. Because the surface form of each gesture-related movement will differ depending on context, the gestures themselves can be considered abstract, although they are not symbolic because they contain intrinsic specifications of quantitative information.

The AP/TD view contrasts with traditional approaches to phonology and phonetics, in which phonological representations are symbolic, and therefore do not define quantitative aspects of articulatory movements or their timing. That is, in traditional approaches, the sequence of symbols /bæt/ cannot be

considered a recipe for generating the quantitative aspects of movements involved in producing the word *bat*. This is because, among other things, the symbols /bæt/ do not specify the movement times, the exact degree of lip compression for the stop closures etc. In contrast, in Articulatory Phonology, phonological representations (i.e. equations of movement and their parameter values) *do* determine aspects of the way constrictions are formed over time. An oft-cited advantage of this approach is that it does not involve translating from one type of representation (categorical symbolic mental representation), to another (representation for specific phonetic form). This advantage is viewed as critically important, because such translation into context-specific variants has been argued to destroy information about the contrastive phonological categories of speech (Fowler, Rubin, Remez and Turvey 1980). The rest of this chapter first presents a general overview of the Task Dynamic approach to motor control (2.3), followed by an introduction to speech motor control in AP/TD (2.4). There follows a detailed discussion of AP/TD mechanisms that relate most closely to speech timing (2.5). Finally, it discusses the specific features that distinguish AP/TD from other approaches, and highlights the advantages of the system (2.6).

2.3 Using mass–spring systems to model gestural movement in TD

The AP/TD model generates articulatory trajectories for planned utterances, which can serve as input to an articulatory synthesizer.[1] In the Task Dynamics framework, gestural movement, or movement toward a constriction goal, is modeled as movement toward an equilibrium position in a damped, mass-spring system, i.e. the movement of a mass attached to a spring (Asatryan and Feldman 1965; Turvey 1977; Fowler et al. 1980). That is, the gestural starting position is analogous to the position to which the mass attached to the spring is stretched, and the equilibrium position is the target position that is approached by the mass after releasing the spring. A mass–spring system can be described as an oscillator, because if the spring is stretched and released, it will oscillate around its equilibrium position in the absence of friction

[1] The model is restricted to articulatory trajectories and does not describe muscle contractions. For discussions of issues relating to modeling muscle contractions, see e.g. Asatryan and Feldman's (1965) equilibrium point hypothesis, and Bullock and Grossberg's (1990) FLETE model.

(i.e. when not damped[2]). Because the system is critically damped, the spring *doesn't* oscillate, but rather reaches within a very short distance of the equilibrium position very quickly, and then continues moving asymptotically toward the equilibrium position but never quite reaches it. When an oscillator is damped (either critically damped, or over-damped) so that instead of oscillating it simply approaches a target, it is said to have point-attractor dynamics. When it oscillates freely because of less-stringent damping, it is said to have limit-cycle dynamics. Both types of oscillators are used within the AP/TD framework, but the discussion here will focus here on point-attractor dynamics, leaving the discussion of limit-cycle oscillators until Section 2.5.3.

A key feature of oscillatory systems with point-attractor dynamics is that they will approximate the equilibrium (target) position, regardless of starting position. The use of this tool in the Task Dynamics model provides a way for the same context-independent phonological unit (a gesture) to have different physical instantiations depending on phonetic context (e.g. the starting position defined by the preceding gestural context). This is because a given gestural dimension is always described by the same equation of motion, with the exception that the value for the starting position parameter is dependent on context. Therefore phonological equivalence can be expressed in terms of the equations of motion that define each gesture (apart from the specification of the starting parameter value).

In addition, in simple systems of this type that have linear damping and restoring forces (spring stiffness), gestural movement duration is proportional to the square root of the stiffness of the spring normalized for its mass, and is predicted to be the same for movements of different amplitude. That is, the spring will move back to equilibrium more quickly when stretched further (cf. Cooke 1980; Ostry and Munhall 1985), and more slowly when stretched less far, resulting in equivalent durations for the movements.[3]

The equation of motion that describes mass–spring oscillations, and is used for each dimension of gestural movement (i.e. for dimensions of constriction location and constriction degree) is the following:

$$m\ddot{x} + b\dot{x} + k(x - x_0) = 0$$

It contains one context-dependent parameter, x, which represents the gestural starting position, and four context-independent parameters: m for mass, b for

[2] An informal analogy of damping might be putting one's feet on the ground to stop a swing.

[3] See Sorensen and Gafos (2016) for a recent proposal for a mass–spring dynamical system with a nonlinear restoring force that makes slightly different and more realistic predictions for the relationship between movement amplitude, speed, and time. See Section 2.5.1.3 for further discussion.

damping, k for spring stiffness, and x_0 for target position. The parameter b (the damping coefficient) is set to a value that ensures critical damping, that is, such that the system reaches very close to equilibrium very quickly, then moves asymptotically toward it, without oscillating. The parameter m (mass) is arbitrarily set to 1 in most implementations (although see Šimko and Cummins 2010 for recent work in which this parameter is varied in a principled way, so that the cost of movement can be computed). This equation defines the movement trajectory for the gestural dimension for each point in time.

In the AP/TD model, values of k (spring stiffness) and x_0 (target position) are specific to the definition of one contrastive category vs. another, and therefore form part of the definition of each gesture (Saltzman, Löfqvist, and Mitra 2000). The values for these parameters that have been chosen in implementations of the model have been estimated from kinematic data. The value of the spring stiffness parameter k is identical for both dimensions (location and degree) that characterize a given gesture (Saltzman and Munhall 1989). Differences in k for consonants vs. vowels are implemented in the model as consistent with empirical data. In particular, the spring stiffness of vowels is lower than that of consonants, and as a result the gestural movements for vowels are slower than those for consonants (Saltzman and Munhall 1989).

The target position x_0 is specific for each dimension of gestural movement (i.e. different for constriction location vs. degree). Differences in x_0 across gesture dimensions are to be expected, since these are required for different constriction locations and degrees characteristic of each linguistic category. For example, the configuration for lip closure for a labial stop will have a different constriction location and degree from the configuration appropriate for an alveolar fricative.

Gafos (2006) and Gafos and Beňuš (2006) have proposed that the underlying target position x_0 for a given gesture can change in particular utterance contexts, according to grammatical and extra-grammatical constraints. This proposal was made in order to account for incomplete voicing neutralization in word-final position in German. In German, e.g., *Rad* 'wheel' and *Rat* 'advice' are both pronounced as [ʁat], but in some experimental conditions the two types of words show subtle differences in voicing during closure, as well as differences in vowel, closure, and burst duration, suggestive of the underlying phonological categories (Port and Crawford 1989; Gafos 2006). Gafos (2006) and Gafos and Beňuš (2006) account for this incomplete neutralization behavior by proposing that the glottal aperture target position x_0 for the final consonant in e.g. *Rad* is under the influence of two weighted, competing attractors, one with a value corresponding to the target position for

voiced stops (consistent with lexical contrast), and the other with a value corresponding to the target position for voiceless stops (consistent with the apparent grammatical de-voicing rule or constraint).

This section has discussed how the equations of mass–spring oscillation are used to describe movements in each gestural dimension; the next section describes two additional aspects of the AP/TD system: how gestures control individual articulators and the mechanism for specifying gestural activation.

2.4 Gestural control of individual articulators, and gestural activation

In the AP/TD framework, the movement of individual articulators is not controlled directly, but rather indirectly, via the selection of gestures (i.e. sets of yoked tract variables) and via gestural activation.

First, each tract variable (e.g. constriction location or constriction degree) controls one aspect of the behavior of a group of articulators that are yoked together (i.e. a coordinative structure) (Table 2.1). Second, the tract variables control constriction formation for each gesture; when a gesture is specified for more than one tract variable, the tract variable specifying location and the tract variable defining degree together define the gesture. And third, there is a gestural score for each utterance, which dictates the amount of time each gesture is active (the gestural activation interval), as well as the relative timing/overlap of different gestures (inter-gestural coordination) (Figure 2.1).

Table 2.1 AP/TD tract variables and the model articulator variables that they govern

Tract Variables		Model Articulators
LP	lip protrusion	upper and lower lips
LA	lip aperture	upper and lower lips, jaw
TDCL	tongue dorsum constriction location	tongue body, jaw
TDCD	tongue dorsum constriction degree	tongue body, jaw
LTH	lower tooth height	jaw
TTCL	tongue-tip constriction location	tongue tip, body, jaw
TTCD	tongue-tip constriction degree	tongue tip, body, jaw
TTCO	tongue-tip constriction orientation	tongue tip, body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottal width

Source: Saltzman et al. (2008). Reproduced with permission.

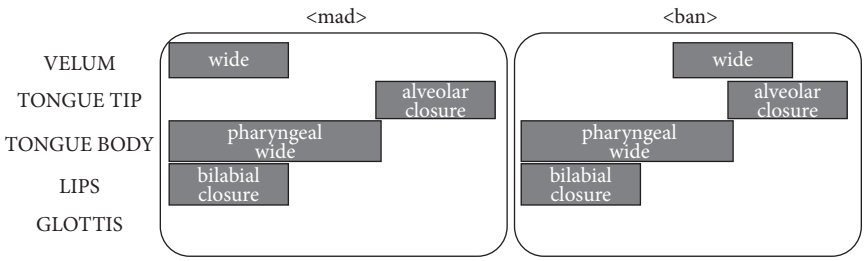

Figure 2.1 Gestural scores for the words *mad* and *ban*, illustrating gestural activation intervals and their relative timing.

Source: Goldstein, Byrd, & Saltzman (2006, p. 226; Figure 7.7). Reproduced with permission from Cambridge University Press © Cambridge University Press 2006

This section discusses the gestural control of individual articulators (Section 2.4.1), as well as the impact on gestural movement of gestural activation, coordination, and the neutral attractor (Section 2.4.2). The details of timing control (both relative timing and the timing of gestural activation) are left until Section 2.5.

2.4.1 Gestural control of individual articulators

Gestures represent movements toward constrictions along a set of relevant dimensions, specified by tract variables. For example, tongue-tip constriction location and degree are the tract variables (dimensions) for tongue-tip gestures; lip aperture and protrusion are the tract variables for lip gestures. Each tract variable, in turn, represents the collective movement of a set of articulators that cooperatively contribute to constriction formation in the dimension specified by the tract variable. These articulator sets are called coordinative structures, or synergies. For example, the separate articulators upper lip, lower lip, and jaw contribute to tract variables for lip aperture and protrusion in lip gestures, while the tongue tip, tongue body, and jaw contribute to tract variables for tongue-tip constriction degree in tongue-tip gestures. Coordinative structures are task- (gesture-)specific. Although there is a default speaker-specific relative contribution of each articulator, the model is configured so that articulators within a coordinative structure can compensate for one another when the need arises. This feature of the model makes it possible for the model to adapt to perturbations. For example, if a load is placed on the jaw during the production of a bilabial sound /p/, the upper lip will

compensate so that lip closure can nevertheless be achieved (e.g. Folkins and Abbs 1975, among many others). In non-perturbed situations, compensation is also seen when a single articulator is involved in the production of multiple overlapping gestures. For example, the jaw is involved in the production of bilabial consonants as well as vowels; if these overlap, and the overlapping vowels are low (low jaw position), the lips will be more involved in the bilabial production than they would be if the overlapping vowel were high (higher jaw position), to compensate for the fact that the jaw is governed by the vowel gesture as well as the consonant gesture. Each articulator will therefore contribute to a gestural tract variable in different proportion depending on context; this type of reorganization is often required in speech.

2.4.2 Gestural activation, overlap in the gestural score, and the neutral attractor

The gestures that specify a particular utterance are organized into a gestural score, which specifies the temporal intervals (gesture activation intervals) during which gestures will be active during the utterance, and patterns of gestural overlap and coordination among gestures. As explained in more detail in Section 2.5, timing at the inter-gestural level is governed by an ensemble of undamped (limit-cycle) planning oscillators, one associated with each gesture (Goldstein et al. 2009), whose frequency, in turn, is influenced by the prosodic level. The prosodic level consists of a set of suprasegmental oscillators (syllable, foot, and phrase, where the foot is defined as a unit extending from one word-level stress to the next; see Saltzman et al. 2008; Nam et al. 2010; Krivokapić 2013).

2.4.2.1 Gestural activation

Gestural activation is schematized by 'boxes' on gestural score diagrams (see Figure 2.2), replaced by slightly different shapes in later versions of the theory. Gestural movement is generated by multiplying the parameters of the mass–spring equation for each tract variable by the gestural activation value at each point in time. This gestural activation value is 0 when activation is off, 1 when it is turned on completely, and an intermediate value if activation is partial; partial activation occurs during on- and off-ramps, as implemented in more recent versions of the model. When it activates a gesture, the activation function also indirectly triggers the movement of the set of articulators controlled by each gesture.

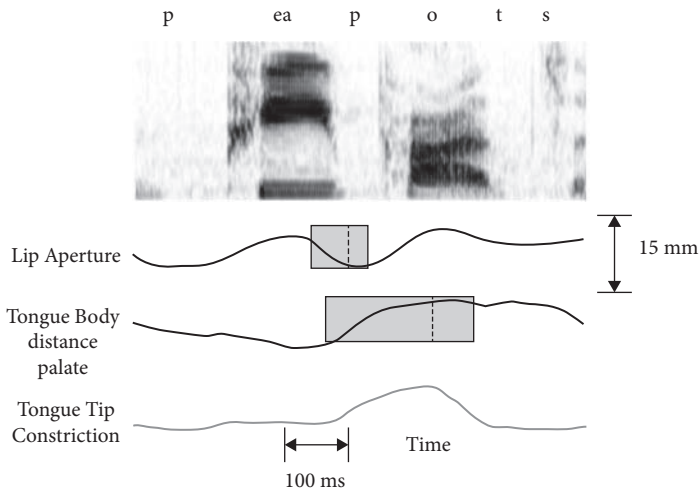

Figure 2.2 Time functions of vocal tract variables, as measured using X-ray microbeam data, for the phrase *pea pots*, showing the in-phase (synchronous within twenty-five ms) coordination of the lip gesture for the /p/ in *pots* and the /a/ gesture for the vowel in *pots*.

Note: Tract variables shown are lip aperture (distance between upper and lower lips), which is controlled for lip closure gestures (/p/ in this example) and tongue-tip constriction degree (distance of the tongue tip from the palate), which is controlled in tongue-tip gestures (/t/ and /s/ in this example).

Also shown is the time function for the distance of the tongue body from the palate, which is small for /i/ and large for /a/, when the tongue is lowered and back into the pharynx. (The actual controlled tract variable for the vowel /a/ is the degree of constriction of the tongue root in pharynx, which cannot be directly measured using a technique that employs transducers on the front of the tongue only. So distance of the tongue body from the palate is used here as a rough index of tongue root constriction degree.)

Boxes delimit the times of presumed active control for the oral constriction gestures for /p/ and /a/. These are determined algorithmically from the velocities of the observed tract variables. The left edge of the box represents gesture onset, the point in time at which the tract variable velocity toward constriction exceeds some threshold value. The right edge of the box represents the gesture release, the point in time at which velocity away from the constricted position exceeds some threshold. The line within the box represents the time at which the constriction target is effectively achieved, defined as the point in time at which the velocity toward constriction drops below the threshold.

Source: Goldstein, Byrd, & Saltzman (2006, p. 230; Figure 7.9). Reproduced with permission from Cambridge University Press.

At a normal, i.e. default, rate of speech, the activation interval is long enough for the gesture to approximate (i.e. reach very close to) its constriction target. However, when the activation interval is shorter than the default (e.g. at faster rates of speech), target undershoot will occur, because the gesture doesn't have enough time to approximate its target. In addition, if the activation interval is longer than the default (because it has been stretched via mechanisms discussed in Sections 2.5.3 and 2.5.4), the articulators will remain

in a quasi-steady state after the target has been approximated, for the remainder of the interval, as the gesture continues to move asymptotically toward the target. Figure 2.2 illustrates lip-aperture movement and tongue-body movement, which remain in a quasi-steady state after the targets have been approximated (dashed lines in grey boxes indicate the point of target approximation).

More detail about the control of activation interval timing is given in Section 2.5.

In sum, gesture activation intervals specify when and how long each gesture will be active; intergestural coordination patterns are specified by the gestural score, as described in the following section.

2.4.2.2 Intergestural coordination

Because the gestural score consists of parallel tiers, one for each gesture, it also specifies patterns of intergestural coordination. That is, the gestural score specifies how gesture activation intervals are timed relative to one another, and thus whether and by how much they overlap. Gestural overlap can have both spatial and temporal consequences. For example, if overlapping gestures make use of the same model articulators (e.g. the jaw is often involved in successive consonant and vowel articulations), then the activity of shared articulators is blended. In such cases the parameter values for the shared articulators will be a combination (i.e. a weighted average) of the parameter values that would have been specified in a non-overlapping configuration. For example, in a VdV sequence, the tongue-tip gesture for /d/ shares tongue-body and jaw articulators with the surrounding vowels, and the activity for these articulators will reflect the combined control of the tongue-tip and vowel gestures. However, tongue-tip activity will be controlled by the tongue-tip gesture alone because it is uniquely involved in the consonant (Öhman 1966; Saltzman and Munhall 1989). Note that if the overlapping gestures share all articulators, target attainment for the overlapped gestures may be compromised. However, AP/TD has mechanisms to ensure target attainment in such circumstances. For example, /g/ in VgV sequences shares both of its oral articulators with the adjacent vowels (i.e. tongue body and jaw). In this situation, the constriction location for /g/ is a result of the combined (overlapping) vocalic and consonantal instructions for the tongue body and jaw, which are shared in the production of /g/ and the surrounding vowels. However, because the blending strength is set to favor constriction degree for consonants over constriction degree for vowels,

the constriction degree target for /g/ can still be reached, with undershoot of the vowel target (Fowler and Saltzman 1993).

Gestural activation intervals and gestural scores specify when the vocal tract is governed by each gesture, but another mechanism is required so that gestural targets can be released. This mechanism, the neutral attractor, is described in the next section.

2.4.2.3 The neutral attractor

When a tract variable is inactive, the articulators which it governs return to their respective neutral positions. These neutral positions are specified by the target of the neutral attractor, which in English is the target configuration for the neutral schwa vowel (Saltzman and Munhall 1989). The neutral attractor governs articulators that aren't governed by active gestures, and thus provides a way of implementing constriction releases. This is because all articulators governed by a gesture will move toward the targets specified by the neutral attractor once the gestural activation interval ends. If gestural activation is partial (e.g. at the beginning or end of a ramped activation interval, Byrd and Saltzman 1998), the vocal tract will be under the simultaneous influence of both the tract variables and the neutral attractor. This is because gestural + neutral attractor activation must always sum to 1 (Byrd and Saltzman 2003).

The neutral attractor yokes together uncoupled, articulator-specific point attractors. Like the tract variables that specify gestural movement, these articulator-specific point attractors are defined by equations that specify movement toward their equilibrium positions (targets). Because the starting position of each articulator-specific point attractor is the point that the articulator has reached at the end of the activation interval for the preceding gesture, the acoustic signal will be influenced by the preceding gesture after it is no longer active (i.e. during the interval governed by the neutral attractor).[4]

The mechanisms that control gestural activation, gestural overlap, and the neutral attractor, described in Sections 2.2, 2.3, and 2.4, provide a general picture of motor control in AP/TD. The following sections provide considerable further detail about timing control mechanisms in particular.

[4] Note that there are thus two types of coarticulation in AP/TD: 1) coarticulation due to gestural overlap, and 2) coarticulation due to the influence of a preceding gesture on the starting position(s) of a articulators governed by an immediately following gesture or neutral attractor.

2.5 Timing Control in AP/TD

In AP/TD, time is included as part of phonological representations (and is therefore *intrinsic* as proposed by Fowler 1977). For how long each gesture shapes the vocal tract (gestural activation intervals), as well as how individual gestures are coordinated with one another (inter-gestural relative timing), are determined by a system of gesture-extrinsic (see Sorensen and Gafos 2016), but phonology-intrinsic, control mechanisms. Speakers do not need to explicitly plan or specify the timing patterns that can be measured from surface acoustics, or surface movement trajectories, because surface timing patterns emerge from the phonological system. Some of the resulting surface timing patterns derive from mass–spring modeling of gestures, whereas others come from the way AP/TD models the control of gestural activation within an utterance, i.e. how it uses prosodic structure and rate control to dictate the amount of time a gesture can shape the vocal tract (its gestural activation interval), and how it models inter-gestural relative timing.

In this section, timing patterns and the way they emerge from control mechanisms are examined for: 1) the timing control of individual gestures (determined by three control mechanisms: gestural stiffness (part of lexical representation), gestural activation (including its rise time), and one gesture-specific, distance-dependent timing adjustment mechanism, discussed in Section 2.5.1, 2) inter-gestural (relative) timing, accomplished through gestural planning oscillator coupling, discussed in Section 2.5.2, and 3) prosodic timing, discussed in Section 2.5.3, which involves two mechanisms: a) trans-gestural timing mechanisms and b) coupled prosodic constituency oscillators, where the coupled prosodic constituency oscillators are also used for global timing control of overall speech rate, discussed in Section 2.5.4. Within this framework, all aspects of timing control are accomplished using oscillators, either critically damped oscillators (for movements toward constrictions and for local timing adjustments), or un-damped, freely oscillating, limit cycle oscillators (for inter-gestural coordination, and prosodic constituent organization).

2.5.1 Timing control of individual gestures

This section discusses three ways in which timing patterns for utterances emerge in AP/TD from mechanisms for the timing control of individual gestures. These three aspects are 1) the surface timing consequences of gestures as mass–spring systems and 2) the amount of time that the vocal tract is

governed by a gesture (the gestural activation interval), as well as 3) mechanisms required to account for differences in movement duration for different distances.

2.5.1.1 Surface timing consequences of gestures as mass–spring systems

As noted above, phonological representations of lexical contrast in AP/TD, i.e. gestures, are spatiotemporal and are modeled as critically damped mass–spring systems. A stretched, critically damped spring will take a predictable amount of time to return to a state very close to equilibrium position (the settling time), and will have a predictable movement time-course during its trajectory. Modeling gestures as critically damped springs in this way has several consequences for timing patterns:

1) *Gestural representations determine gestural settling time*, or the time it would take the gesture to approximate its target, assuming the gesture is fully active and active for long enough. Because each contrastive gesture is described phonologically as a mass–spring system, i.e. as an equation of oscillatory motion, the time it takes to approximate a target is a function of the mass, damping, and stiffness parameters of the equation. In the fully implemented Saltzman et al. (2008) model, damping is always fixed to critical, and mass is also invariant across phonological categories (but see Šimko and Cummins 2010 for an alternative approach). Temporal differences across phonological categories in the current version of AP/TD therefore relate exclusively to k, the stiffness parameter. Vowels are assumed to have lower stiffness than consonants (and consequently longer mass–spring settling times); as a result, vowels have slower movements toward vowel targets as compared to the movements toward consonantal targets.
2) *Movements are produced with a smooth, single-peaked tangential velocity profile*. The point attractor mass–spring dynamics of this model generates the hallmark of practiced, purposeful movements: a smooth, single-peaked, tangential velocity profile, i.e. with a single acceleration and a single deceleration phase. However, as discussed in more detail below, the velocity profiles generated by systems with linear restoring forces (as in the original AP/TD model) have velocity peaks that are much earlier than observed in empirical data. An extra mechanism, i.e. gradual activation interval on- and off-ramps, was therefore added to the original system to create more realistic velocity profiles. A more recent proposal with a nonlinear restoring force (Sorensen and Gafos

2016) can generate more realistic timing of the velocity peak without the extra mechanism.

3) *Mass-spring dynamics predicts that movement peak velocity will be faster for movements of longer distances*. In mass–spring systems with a given stiffness specification, the peak velocity is faster if the movement distance is longer. In fact, in mass–spring systems with linear restoring forces (Saltzman et al. 2008), the ratio of movement distance to movement peak velocity is proportional to mass–spring settling time, resulting in equal durations for movements of longer distance as compared to movements of shorter distance (cf. Ostry and Munhall 1985). Put another way, gestural movements of different distances are predicted to have the same duration if their stiffness specifications are the same. Observations of speech data do indeed show that peak velocity is faster for longer-distance movements as compared to shorter-distance movements, as this mechanism predicts. However, durations for movements of different distances nevertheless do differ (e.g. Ostry, Keller, and Parush 1983) (sometimes described as 'the farther, the longer' phenomena), and therefore require additional mechanisms to account for them (see Section 2.5.1.3). An alternative in the form of mass–spring dynamics with a nonlinear restoring force has been suggested by Sorensen and Gafos (2016). See also Chapter 8 for an alternative explanation in the Optimal Control Theory framework.

2.5.1.2 The amount of time the vocal tract is governed by a gesture: the Gestural Activation Interval

Activation intervals determine the amount of time for which the vocal tract is intended to be shaped by the movements of a given gesture. These intervals are controlled by a hierarchy of coupled planning and suprasegmental oscillators. In recent developments of AP/TD, gestural activation intervals are not specified in terms of milliseconds (or other units that correlate with solar time). Instead, each gestural activation interval corresponds to a fixed proportion of a gestural planning oscillator period, and the oscillation frequency of the planning+suprasegmental oscillator ensemble determines the gestural activation interval. The oscillation frequency of this planning+suprasegmental ensemble can be varied according to desired speech rate, and can be adjusted at appropriate prosodic positions (e.g. boundaries and prominences), in order to stretch the activation intervals at these positions. As discussed, the amount of time required to approximate the target will be dictated by gestural stiffness. At a normal, default, rate of speech, the activation interval for each gesture is