

The dynamics of phonological planning

by

Kevin D. Roon

A dissertation submitted in partial fulfillment

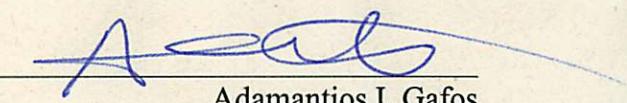
of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics

New York University

January, 2013



Adamantios I. Gafos

© Kevin D. Roon

All Rights Reserved, 2013

## **DEDICATION**

*To Simon*

## **ACKNOWLEDGMENTS**

I am grateful and honored to have had Adamantios Gafos as my advisor for my dissertation. I am impressed with the persistence he showed and deft hand he used in working with me to find the topic that I ultimately have wound up addressing here. It is hard to communicate how edifying and gratifying it has been to work through every aspect of this research with someone so exacting in theory, detail, thinking, and writing. If this dissertation marks the start of a research program with Diamandis rather than the culmination of one, I will be doubly fortunate.

I am also deeply indebted to the rest of my dissertation committee. I thank Lisa Davidson and Maria Gouskova not only for their extensive feedback on this dissertation, but also for leading excellent courses and seminars, for providing me with very rewarding research opportunities and invaluable training, and for being extremely supportive during my graduate career. Thanks to both of them I have also been able to keep my interest in Russian alive and productive, independently from this dissertation. I thank Ioana Chitoran for her steadfast interest in my work even as it changed, for always providing thoughtful perspective on it, and for her professional advice and help. I thank Alec Marantz for his clear-headed guidance, and especially for overseeing my foray into working with response time data in my first qualifying paper.

I am sure that this dissertation and my graduate school experience as a whole would have suffered immeasurably had I not shared an office with Jason Shaw and Tuuli Morrill for several years. I am grateful for the unbelievably generous amount of time that Jason spent discussing this work with me, digging into the details of it all and maintaining a level of interest and enthusiasm that I found inspiring. This work has also benefitted greatly from discussions with Tuuli, who was always there to provide invaluable perspective and support. Both of them have become great friends as well as wonderful colleagues, for which I feel very privileged.

I also thank my fellow grad students Rahul Balusu and Amanda Dye (in addition to Jason) who took part in the seminars at NYU that helped shape and refine this research project.

I would like to thank several people at Haskins Laboratories for their help with various aspects of this research. I thank Carol Fowler, Louis Goldstein, and Doug Whalen for their interest in my research and for helpful discussions about it as it developed. I would like to thank others at Haskins for practical advice and assistance in addition to helpful discussions. Bruno Galantucci provided invaluable advice on designing my experiments, as well as scripts and other files for running them. Tine Mooshammer helped me sort through the statistical analyses of my data. Khalil Iskarous was of great help in automating certain aspects of the data-labeling process.

I also thank Kathy Rastle for advice on running experiments, Chris Kirov for help with implementation of MATLAB coding and discussing the nitty-gritty of DFT,

Anna Greenwood for help with organizing my experimental materials for use beyond the work presented here, and Jon Brennan for lots of help with statistics and R.

I am indebted to the audiences at LabPhon 12 (Albuquerque) and 13 (Stuttgart), NECPhon 3 (MIT) and 4 (UMASS), LSA 2011 (Pittsburgh), Dartmouth College, the International Workshop on Language Production 2012 hosted by NYU, and the Universität Potsdam for feedback on the work as it progressed.

I am grateful to the National Science Foundation for the Dissertation Improvement Grant that helped support this research, and to the anonymous reviewers who voiced confidence in this research. This material is based on work supported by the National Science Foundation under Grand No. 0951831. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

I can think of no more appropriate place than my doctoral dissertation to thank my parents for their commitment to providing me with the best possible educational opportunities throughout my life, and the freedom to pursue whatever path I chose.

Last and certainly not least I thank Simon for his encouragement and support, our wonderful life, and for listening to far more about phonetics, phonology, and the speech perception-production link than any non-linguist should ever have to.

## ABSTRACT

This dissertation proposes a dynamical computational model of the timecourse of phonological parameter setting. In the model, phonological representations embrace phonetic detail, with phonetic parameters represented as activation fields that evolve over time and determine the specific parameter settings of a planned utterance.

Existing models of speech production assign little or no role to phonological features, and theories of phonological features lack the notion of the timecourse of how those features get set. One benefit of the model presented here is that it provides a formal link between speech perception and production, which has been notably missing in the literature despite a longstanding debate on the topic (cf. Diehl, Lotto, & Holt, 2004).

This dissertation capitalizes on the convergence of novel experimental and computational results to identify specific requirements of any model of the perception-production link, including a role for representations at the level of phonological features and the computational principles of both excitation and inhibition.

Another benefit of this dynamical model is that it enables establishing formal links between phonological processes and response time data. The model accounts for response times in a task in which speakers hear distractors as they are preparing to produce utterances. Previous studies using this task (e.g., Galantucci, Fowler, & Goldstein 2009) have found that subjects produce an utterance more quickly when they perceive a distractor that is identical to a response being planned than when it is

different. The perception-production link is modeled here as the influence of a perceived distractor on the process of setting the phonological production parameters of a required utterance. Response time modulations are due to the effects of combining (in)compatible inputs to this planning process. The model predicts gradient effects on response times based on the degree of similarity between a distractor and a response, with responses being quickest when they are identical, slower when they differ on one parameter (voicing or articulator), and slower still when they differ on more than one parameter. These predictions are confirmed in two experiments that provide the first clear evidence of perceptuo-motor effects of voicing and articulator.

## TABLE OF CONTENTS

|  |     |
|--|-----|
| DEDICATION   | iii |
| ACKNOWLEDGMENTS                                    | iv  |
| ABSTRACT   | vii |
| LIST OF FIGURES                                    | xv  |
| LIST OF TABLES                                     | xix |
| LIST OF APPENDICES                                 | xxi |
| CHAPTER 1: Introduction                            | 1   |
| 1.1. Claims  | 1   |
| 1.2. Organization                                  | 1   |
| CHAPTER 2: The timecourse of phonological planning | 4   |
| 2.1. Introduction                                  | 4   |
| 2.2. Phonological planning                         | 4   |
| 2.3. Perception-production link                    | 7   |
| 2.3.1. Evidence from phonetic modulations          | 8   |
| 2.3.2. Evidence from response time modulations     | 10  |
| 2.3.3. Evidence from neuroimaging                  | 12  |
| 2.4. Discussion                                    | 14  |
| 2.4.1. Is identity special?                        | 15  |
| 2.4.2. Speech perception debate                    | 24  |

|   |    |
|---|----|
| 2.5. Conclusion                               | 25 |
| CHAPTER 3: Experiments                        | 27 |
| 3.1. Introduction                             | 27 |
| 3.2. Voicing RT experiment (Experiment 1)     | 29 |
| 3.2.1. Methods                                | 29 |
| 3.2.1.1. Subjects                             | 29 |
| 3.2.1.2. Procedure                            | 30 |
| 3.2.1.3. Equipment                            | 33 |
| 3.2.1.4. Stimuli                              | 35 |
| 3.2.2. Results                                | 37 |
| 3.2.2.1. Statistics                           | 43 |
| 3.2.2.2. Results of statistical models        | 49 |
| 3.2.2.2.1. Control effects                    | 49 |
| 3.2.2.2.2. Experimental effects               | 57 |
| 3.2.3. Discussion                             | 64 |
| 3.3. Articulator RT experiment (Experiment 2) | 66 |
| 3.3.1. Methods                                | 66 |
| 3.3.1.1. Subjects                             | 66 |
| 3.3.1.2. Procedure                            | 67 |
| 3.3.1.3. Stimuli                              | 68 |

|  |     |
|--|-----|
| 3.3.2. Results   | 69  |
| 3.3.2.1. Control effects                                   | 73  |
| 3.3.2.2. Experimental effects                              | 74  |
| 3.3.3. Discussion  | 78  |
| 3.4. General Discussion                                    | 79  |
| 3.4.1. Methodological observations                         | 81  |
| 3.4.2. Comparison with previous studies                    | 85  |
| 3.5. Conclusion  | 90  |
| CHAPTER 4: Motivating excitation and inhibition            | 91  |
| 4.1. Introduction  | 91  |
| 4.2. Experimental evidence                                 | 91  |
| 4.2.1. Linguistic and non-linguistic influences on RTs     | 92  |
| 4.2.2. Excitation and inhibition                           | 93  |
| 4.3. Do “common codes” excite or inhibit?                  | 96  |
| 4.3.1. Accessing a common code                             | 97  |
| 4.3.2. Repeated phonological planning                      | 101 |
| 4.3.3. Revised phonological planning                       | 104 |
| 4.4. Conclusion  | 111 |
| CHAPTER 5: Dynamic Field Theory and phoneme classification | 112 |
| 5.1. Introduction  | 112 |

|   |     |
|---|-----|
| 5.2. Phoneme classification                     | 112 |
| 5.3. Model and simulations                      | 115 |
| 5.3.1. Dynamic Field Theory                     | 115 |
| 5.3.2. Model of the phoneme classification task | 117 |
| 5.3.2.1. Activation fields                      | 118 |
| 5.3.2.2. Model dynamics                         | 119 |
| 5.3.2.3. Pre-shapes                             | 122 |
| 5.3.2.4. Stimulus input                         | 125 |
| 5.3.2.5. Interaction term                       | 126 |
| 5.3.2.6. Noise                                  | 129 |
| 5.3.2.7. Summary                                | 129 |
| 5.3.3. Experiment simulations                   | 130 |
| 5.3.3.1. Good-exemplar stimuli                  | 134 |
| 5.3.3.2. Marginal-exemplar stimuli              | 137 |
| 5.3.3.3. Ambiguous-exemplar stimuli             | 139 |
| 5.4. Discussion                                 | 140 |
| 5.4.1. Results of the model                     | 141 |
| 5.4.2. Pre-shapes                               | 142 |
| 5.4.3. Comments on the dynamics of the model    | 147 |
| 5.5. Summary and conclusions                    | 149 |

|   |     |
|---|-----|
| CHAPTER 6: A dynamical model of the timecourse of phonological planning | 151 |
| 6.1. Introduction   | 151 |
| 6.2. Model components   | 153 |
| 6.2.1. Planning fields  | 154 |
| 6.2.2. Input  | 159 |
| 6.2.3. Monitor and Implementation                                       | 163 |
| 6.3. Model dynamics   | 163 |
| 6.3.1. Pre-shapes   | 166 |
| 6.3.2. No inhibition: Tone condition                                    | 170 |
| 6.3.3. Cross-field inhibition: Congruent condition, Experiment 1        | 175 |
| 6.3.4. Within-field inhibition: Congruent condition, Experiment 2       | 179 |
| 6.3.5. Cross- and within-field inhibition: Incongruent condition        | 182 |
| 6.3.6. Reinforcing input: the Identity condition                        | 188 |
| 6.4. Simulations  | 192 |
| 6.5. Model variable values  | 196 |
| 6.6. Discussion   | 199 |
| 6.6.1. Determining variable values                                      | 200 |
| 6.6.2. Generality of the model  | 205 |
| 6.6.3. Incongruity effects  | 207 |
| 6.6.4. Unknown voicing vs. unknown articulator                          | 212 |

|  |     |
|--|-----|
| 6.6.5. Additional predictions                      | 216 |
| 6.7. Conclusion                                    | 218 |
| CHAPTER 7: Discussion and conclusion               | 220 |
| 7.1. Summary                                       | 220 |
| 7.2. Theoretical implications                      | 221 |
| 7.2.1. Perception-production link in speech        | 221 |
| 7.2.2. Theories of speech production               | 224 |
| 7.2.3. Identity                                    | 225 |
| 7.3. Future research                               | 226 |
| 7.3.1. Testing additional predictions of the model | 226 |
| 7.3.2. “Articulator” effects                       | 227 |
| 7.3.3. Phonetic modulation                         | 230 |
| 7.3.4. Remaining questions                         | 234 |
| 7.4. Advantages of dynamical modeling              | 237 |
| 7.5. Conclusion                                    | 239 |
| APPENDICES   | 242 |
| REFERENCES   | 266 |

## LIST OF FIGURES

|           |  |    |
|-----------|--|----|
| Figure 1  | Comparing temporal and spatial properties of a) <i>ba</i> , b) <i>da</i> , c) <i>pa</i> , d) <i>ta</i> .   | 14 |
| Figure 2  | Timeline of one trial.   | 31 |
| Figure 3  | Schematic of the hardware setup for the experiments.   | 34 |
| Figure 4  | An example of the data labeling for one token of a subject replying <i>ta</i> .  | 40 |
| Figure 5  | Mean RTs for the voicing experiment (Experiment 2).  | 42 |
| Figure 6  | Mean RTs in ms for Experiment 1, by block.   | 43 |
| Figure 7  | A) The non-transformed response-time data for all subjects and conditions was positive skewed. B) The same data log-transformed was closer to a normal distribution.                                       | 46 |
| Figure 8  | Log response times over the course of the experiment (both blocks) by subject.   | 51 |
| Figure 9  | Comparison of the intercept and slope assigned when centered Trial is modeled only as a fixed effect (dotted line) with the interaction of Trial and Subject also modeled as a random effect (solid line). | 54 |
| Figure 10 | Mean RTs for the articulator experiment (Experiment 2).  | 71 |
| Figure 11 | Mean RTs in ms for Experiment 2, by block.   | 72 |
| Figure 12 | Results from experiment 2 of Galantucci, Fowler, and Goldstein (2009).   | 85 |

|  |     |
|--|-----|
| Figure 13 Schematic, qualitative comparison of distractor conditions based on Experiment 2 of Galantucci, Fowler, and Goldstein (2009) and the results from the present experiment 2.                                  | 88  |
| Figure 14 Results of the interference task with arrow presses of Müsseler (1995), as reported in Prinz (1997).   | 99  |
| Figure 15 Interaction-activation model of syllable production from Meyer and Gordon (1985, p. 20, Figure 2), showing a trial where the primary response was <i>ut-ub</i> and the secondary response was <i>ub-ut</i> . | 108 |
| Figure 16 Previous results from phoneme-classification experiments.  | 113 |
| Figure 17 Components of the model of the phoneme-classification task.  | 118 |
| Figure 18 Evolution of two fields with input at 40 ms VOT differing only in the value of $\tau$ : (A) $\tau = 80$ and (B) $\tau = 240$ .   | 120 |
| Figure 19 Activation field returning to resting level.   | 121 |
| Figure 20 Pre-shape inputs.  | 123 |
| Figure 21 Inputs of good-exemplar stimuli for <i>ta</i> with a VOT of 40 ms (solid red line) and <i>da</i> with a VOT of 0 ms (dashed blue line).  | 125 |
| Figure 22 The interaction term $w(x)$ , showing the values of (4) used in model.   | 126 |
| Figure 23 Maximum activation levels of the TA and DA fields with the same pre-shape input only (i.e., without any stimulus input) with three different values for $\beta$ : 1.5 (A), 0.5 (B), and 0 (C).               | 128 |

|           |   |     |
|-----------|---|-----|
| Figure 24 | The results of the simulations of 500 trials of the phoneme classification task, with 100 trials for each stimulus VOT value.   | 131 |
| Figure 25 | Shows the evolution of the activation fields for simulations of three categories of stimuli in the phoneme-classification task. | 133 |
| Figure 26 | Histograms of VOTs of <i>da</i> (A) and <i>ta</i> (B) productions of all speakers from the experiments reported in Chapter 3.   | 143 |
| Figure 27 | Pre-shapes modeled as normal distributions as defined in (3).   | 144 |
| Figure 28 | Model simulations including a <i>ta</i> stimulus with VOT = 90 ms.  | 145 |
| Figure 29 | Components of the dynamical computational model of phonological planning in the response-distractor task.                       | 153 |
| Figure 30 | Planning fields for <i>ta</i> .   | 156 |
| Figure 31 | Pre-shapes without other input.   | 167 |
| Figure 32 | Evolution of the Voicing field for a simulated trial with no within-field inhibition.   | 169 |
| Figure 33 | Activation field evolutions in the Tone condition of experiment 2.  | 171 |
| Figure 34 | Activation field evolutions in the Tone condition of experiment 1.  | 174 |
| Figure 35 | Comparison of activation field evolutions showing cross-field inhibition.   | 177 |

|  |     |
|--|-----|
| Figure 36 Comparison of activation field evolutions showing within-field evolution from a simulated trial from the Congruent condition of the <i>ta-ka</i> block of experiment 2, where the required response was <i>ta</i> , the distractor was <i>da</i> , and the SOA was 100 time steps. | 181 |
| Figure 37 Comparison of activation field evolutions in the Incongruent conditions of experiments 1 and 2.  | 184 |
| Figure 38 Comparison of the activation level evolutions in the Identity case.  | 190 |
| Figure 39 Results from model simulations of the response-distractor task.  | 194 |
| Figure 40 Sources of inhibition in the Congruent and Incongruent conditions in the two experiments, assuming that within-field inhibition imposes a slightly smaller slow-down on RTs than cross-field inhibition, as indicated by the width of the boxes.                                   | 208 |
| Figure 41 Simulation of a modified version of experiment 1, where the potential responses are <i>ta</i> or <i>ka</i> , and response is <i>ta</i> , the distractor is <i>ba</i> , and the SOA is 100 time steps.  | 211 |
| Figure 42 RTs by response syllable.  | 215 |

## LIST OF TABLES

|            |  |    |
|------------|--|----|
| Table I    | Response-distractor pairs for the voicing experiment.  | 32 |
| Table II   | Properties of the distractor stimuli in the voicing experiment.  | 35 |
| Table III  | Results of Kolmogorov-Smirnov tests for normality on the non-transformed RTs in milliseconds and log-transformed response-time data.                             | 47 |
| Table IV   | Control random and fixed effects on RTs included in the model.   | 52 |
| Table V    | Results of a linear mixed-effects model of control effects on Log RT in the voice experiment.  | 56 |
| Table VI   | Experimental effects on RT added into the model.   | 58 |
| Table VII  | Results of a linear mixed-effects model including the experimental effects on Log RT in the voice experiment (Experiment 1).                                     | 60 |
| Table VIII | ANOVA comparison of the model with control effects only with the model that included the experimental effects of Distractor congruency.                          | 61 |
| Table IX   | Results of a linear mixed-effects model including linguistic effects on Log RT in the voice experiment, without subject-specific slopes for Trial or Distractor. | 62 |

|             |  |     |
|-------------|--|-----|
| Table X     | Results of Markov chain Monte Carlo sampling (with 10000 samples) for the voice experiment.  | 63  |
| Table XI    | Response-distractor pairs for the articulator experiment.  | 67  |
| Table XII   | Properties of the distractor stimuli in the articulator experiment.  | 69  |
| Table XIII  | Results of a linear mixed-effects model of control effects on Log RT in the articulator experiment.                                  | 74  |
| Table XIV   | Results of a linear mixed-effects model including experimental effects on Log RT in the articulator experiment.                      | 75  |
| Table XV    | ANOVA comparison of the model with only control effects with the model that included the experimental Distractor congruency effects. | 77  |
| Table XVI   | Results of Markov chain Monte Carlo sampling (with 10000 samples) for the articulator experiment.                                    | 78  |
| Table XVII  | Hand-measured RTs vs. voice key.   | 83  |
| Table XVIII | Comparison of distractors and responses in Galantucci, Fowler, and Goldstein (2009, “GFG”)’s experiment 2 and present experiments.   | 86  |
| Table XIX   | Stimuli to test predictions of the model regarding articulator vs. tract-variable effects.   | 229 |

## **LIST OF APPENDICES**

|            |  |     |
|------------|--|-----|
| APPENDIX A | MATLAB code (categorize_trial.m) implementing the simulation of one trial of the phoneme-classification task using the computational model described in Chapter 5.                         | 242 |
| APPENDIX B | MATLAB code (categorize_exp.m) for implementing simulations of the phoneme-classification experiment, i.e., multiple trials for any number of stimuli, as described in Chapter 5.          | 248 |
| APPENDIX C | MATLAB code (dft_resp_distr.m) implementing the simulation of one trial of the response-distractor task using the computational model described in Chapter 6.                              | 250 |
| APPENDIX D | MATLAB code (simulate_exps.m) implementing the simulation of the response-distractor experiment, i.e., multiple trials and conditions, using the computational model defined in Chapter 6. | 263 |

## **CHAPTER 1: INTRODUCTION**

### **1.1. Claims**

In this dissertation I make the case that there is benefit to be gained by developing a formal, computational model of the timecourse of the process by which phonological parameter values are set during speech production. The development of the model has two main motivations. First, while models from the literature on speech production explicitly include the timecourse of the processes involved, there is little if any consensus on to what extent representations at the level of phonological features are involved. Some models assign no role to them at all. On the other hand, no model of phonological representation addresses the issue of how values for those representations are set in real time during production of an utterance. Second, despite longstanding debate on the nature of the link between speech perception and speech production, no formal models of this link have been proposed. In light of new experimental evidence presented in this dissertation, I claim that developing a dynamical model of phonological planning is in fact warranted, and that such a model provides a means to formalize the link between speech perception and production.

### **1.2. Organization**

Chapter 2 surveys theories of speech production and phonological representation, highlighting the gap noted above. The gap is explained in large part

because empirical evidence in favor of assigning a role to phonological features in production and/or the perception-production link has been conspicuously lacking. Chapter 3 addresses this question by presenting data from two new experiments that establish independent roles for voicing and articulator in the perception-production link. Chapter 4 discusses the experimental results from Chapter 3 and compares them with results from other studies, identifying necessary properties for any formal account of the process of speech production and its link with speech perception. First, a role for representations at the level of phonological features is needed. Second, any such model must incorporate the computational principles of excitation and inhibition. Chapter 5 introduces and explains the specific computational framework that is used in Chapter 6 to develop a model of the task used in the experiments in Chapter 3. Chapter 5 also shows that this framework is capable of accounting for both gradient response time data and categorical classification data from a widely used phoneme classification task. Chapter 6 presents a formal computational model of phonological planning using the experimental task from Chapter 3. The model formalizes the link between perception and production as the phonological parameter values of a perceived utterance serving obligatorily as input to the ongoing phonological planning of an utterance. The model provides an account of the experimental results from Chapter 3, the results of previous studies, and an unexpected difference in response times across the two experiments in Chapter 3. Chapter 7 concludes with a discussion of the theoretical implications of the results of both the experiments and the

computational model, and indicates future research directions based on questions that they raise.

## **CHAPTER 2: THE TIMECOURSE OF PHONOLOGICAL PLANNING**

### **2.1. Introduction**

This dissertation presents a dynamical model of the timecourse of phonological planning. An immediate benefit of the model is that it provides a way to formalize the link between speech perception and production. Despite a longstanding debate on perception-production link (see Diehl, Lotto, & Holt, 2004; Galantucci, Fowler, & Turvey, 2006), a specific proposal formalizing this link has been absent in the literature. Section 2.2 addresses what is meant by phonological planning, and shows that an adequate model of that process has been lacking. Section 2.3 reviews various experimental results that have been presented as evidence for the link between speech perception and production. Section 2.4 addresses further studies that raise the question of what properties of speech are involved in the perception-production link, specifically, whether there is reason to expect that properties smaller than segmental identity, e.g., properties at the level of phonological features, might play a role in this link. Section 2.5 concludes.

### **2.2. Phonological planning**

Detailed theories exist of the dynamics by which speech movements are physically implemented. In Guenther's DIVA model (Guenther, 1995; Guenther, Ghosh, & Tourville, 2006; Guenther, Hampson, & Johnson, 1998), movements of

speech articulators are guided to achieve acoustic speech targets. In the DIVA model, the timecourse of the process of implementation is crucial, as real-time adjustments to articulator directions and velocities are made so as to achieve the required acoustic output as closely as possible. The timecourse of movements is also a core component of the Task Dynamics Model of Saltzman and Munhall (1989), where articulator movements are controlled to achieve articulatory goals. In the Task Dynamics Model, movements of sets of articulators are coordinated to effect specific constriction locations of the various primary speech articulators. A core difference between DIVA and Task Dynamics lies in whether speech goals are acoustic or articulatory in nature. This question is not of primary concern to the questions asked in this dissertation (though see, e.g., Boersma, 1998, for a thorough discussion). The important point is that these two theories have in common an explicit incorporation of timecourse in their models. A review of the existing theories that set the parameters that are implemented by a model like DIVA or the Task Dynamics Model shows either that little to no importance is given to the setting of phonological features, or that no attention is paid to the timecourse by which such setting takes place.

Models of speech production (see Levelt, 1999; Roelofs, 2000, for reviews) encompass processes of both word/lemma selection and word-form encoding (Dell & O'Seaghda, 1992; Levelt, 1992; MacKay, 1987), the latter including various types of phonological encoding. Two major models that have been computationally implemented are the spreading activation model of Dell and colleagues (Dell, 1986,

1988; Dell, Juliano, & Govindjee, 1993; Dell & O'Seaghda, 1992) and the WEAVER++ model (Levelt, Roelofs, & Meyer, 1999; Roelofs, 1997, 2000). Both models include the timecourse of all component processes as a critical dimension of their implementation, and both models assume that their output serves as the input to some other component that physically implements the requisite speech articulations, like those outlined above. The relevant distinction between the two models for the present discussion is the role assigned to phonological features. In WEAVER++, the phonemes of a word are used to select the appropriate stored syllable from a syllabary. This syllable has associated articulatory plans, which are then sent to production. The phonological encoding of WEAVER++ does not include any role for features. In contrast, Dell's model does assign a specific role for features in order to account for speech errors that are attributed to the application of phonological rules/constraints. However, the role that is assigned to phonological features in Dell's model is very limited, providing the model with "only [...] a mechanism for sensitivity to phonotactics constraints, the consonant and vowel categories, and syllabic constituency" (Dell et al., 1993, p. 180).

On the other hand, theories of phonological representation, whether binary phonological features of Chomsky and Halle (1968) or gestural scores of Articulatory Phonology (Browman & Goldstein, 1986; 1989, et seq.), do not include a dynamical component for addressing how their values are established during speech production. In Articulatory Phonology, for example, the Linguistic Gestural Model is the

component that generates a gestural score (Browman & Goldstein, 1990). This gestural score in turn serves as input to the Task Dynamics Model. The timecourse by which this gestural score is generated is not formalized in Articulatory Phonology, even though it stands to reason that this process must have a timecourse: no process can take no time.

Bohland, Bullock, and Guenther (2009) propose a model (Gradient Order DIVA, or GODIVA) that does address the timecourse of phonological encoding for strings of speech sounds that are then implemented by DIVA. However, GODIVA is similar to WEAVER++ in that the phonological representations that it has implemented are only segments and syllables. The authors note that although they do not include representation for any types of phonological features, the model was designed such that it could be expanded to include feature-level representations, should that be warranted.

I argue that such an expansion is warranted, motivated by experimental evidence presented in Chapter 3. These experiments explore whether and how representations at the level of phonological features play a role in the way the speech perception and speech production systems interact.

### **2.3. Perception-production link**

Several studies (Galantucci, Fowler, & Goldstein, 2009; Gordon & Meyer, 1984; Kerzel & Bekkering, 2000; Nielsen, 2007; Yuen, Brysbaert, Davis, & Rastle,

2010) have shown that the phonetic detail of the speech speakers produce, and how quickly they produce it, can be modulated systematically and involuntarily by various stimuli they perceive while speaking. These studies indicate that during speech perception there is an interaction with the speech production system that seems inescapable, at least for normal speaker-hearers. Phonetic and response-time modulations that are attributed to this interaction are referred to as “perceptuo-motor effects” (Galantucci et al., 2009; Kerzel & Bekkering, 2000).

There are three types of experimental evidence for the link between speech perception and production that are reviewed below. The first is modulation of phonetic detail of produced utterances as a result of perceived utterances. The second is modulation of response times (RTs) of produced utterances based on utterances perceived during production planning. Third, there is neuroimaging evidence in support of a strong link between speech perception and production.

### **2.3.1. Evidence from phonetic modulations**

There is ample experimental evidence showing that the phonetic characteristics of utterances that subjects produce can be influenced by utterances they perceive. Studies by Fowler and colleagues show that speakers make phonetic adjustments to Voice Onset Time ("VOT", Lisker & Abramson, 1964) based on perception. Sancier and Fowler (1997) showed that the VOTs of voiceless stops produced by a Brazilian bilingual Portuguese-English speaker were affected by ambient language environment.

They found that the speaker's VOTs in Portuguese were longer after an extended period in the US, and shorter after exposure to Portuguese in Brazil. Upon returning to the US, this exposure to Portuguese in turn significantly reduced her VOTs in English. Fowler, Brown, Sabadini, and Weihing (2003) found that subjects' VOTs were affected by stimulus VOTs, which they manipulated in a shadowing task where speakers repeated auditory stimuli. Subjects produced significantly longer VOTs when they heard longer VOTs in the stimulus. Shockley, Sabadini, and Fowler (2004) found that speakers' VOTs for word-initial voiceless stops increased in duration from one recording session to another when in between sessions they heard the same words they were producing whose VOTs had been digitally extended.

Nielsen (2007) also shows that VOTs can be modulated by input, and additionally that speakers generalize VOT differences introduced by one phoneme to other phonemes. Subjects read lists of English words aloud. The words of interest had simplex onsets of either *k* or *p*. Subjects then heard a recording of two repetitions of a subset of the words containing only some of the *p*-initial words and none of the *k*-initial words. VOTs of *p*-initial words had been lengthened to about 113 ms. Subjects then re-read the list after hearing the modified stimuli. Their VOTs were significantly longer when read after hearing the stimuli. The effect was not limited to the subset of words that were heard in the stimulus list. Increased VOTs were found for *p*-initial words that were not heard in the stimuli, and for *k*-initial words, none of which were heard in the stimuli.

Yuen et al. (2010) report an electro-palatography study where subjects produced *k*- and *s*-initial utterances while hearing compatible (*k*- or *s*-initial, respectively) or incompatible (*t*-initial) distractors that immediately preceded the orthographic cue of the response. They report increased alveolar closure for all responses when subjects heard the incompatible distractor compared to when they heard a compatible distractor. They attribute this additional degree of alveolar closure in the *s* and *k* productions to the involuntary activation of motor commands for tongue-tip closure as a result of perceiving a *t*-initial utterance.

### **2.3.2. Evidence from response time modulations**

This sub-section presents two studies that have used evidence from the RTs of speakers' utterances as evidence for the link between perception and production. The assumption of the authors of these studies is that a perceived stimulus increases the activation of the "speech motor system" needed to produce the perceived stimulus (Galantucci et al., 2009). Due to this increased activation, speakers produce an utterance more quickly if they have recently perceived some stimulus that activates the same motor plans as the utterance they need to produce.

Kerzel and Bekkering (2000) had subjects learn visual stimulus-spoken syllable pairings (e.g., if you see && say *ba*, if you see ## say *da*). Responses were always either *ba* or *da*. In their experiments, visual distractor stimuli were presented while subjects were preparing to produce the *ba* or *da* syllables. Distractors were silent

videos of a speaker mouthing a syllable. Videos were either congruent or incongruent with the subject's response. In congruent videos, a speaker was shown mouthing the same syllable that the subject had to produce on a given trial, e.g., mouthing *ba* when the subject was supposed to say *ba*. In incongruent videos, a speaker was shown mouthing the syllable that the subject was not supposed to produce, e.g., mouthing *da* when the subject was supposed to say *ba*. Subjects were told to ignore the distractors. The timing of the visual distractors relative to the cue stimuli was manipulated with distractors appearing before, synchronous with, or after the cue stimuli (that is, they had varying stimulus onset asynchronies, hereafter "SOAs"). Kerzel and Bekkering (2000) found significantly shorter RTs in this response-distractor task when the speakers' responses were accompanied by a congruent distractor than when accompanied by an incongruent distractor. The authors attributed these shorter RTs in the congruent condition to perceptuo-motor effects of articulator<sup>1</sup>, though as will be discussed below, there is some reason to question this interpretation.

Galantucci et al. (2009) report experiments similar to those of Kerzel and Bekkering (2000), but instead of videos they used auditory distractors that were either identical to or different from the speakers' responses. They found (in their experiment 2) that RTs were significantly shorter when the distractor was the same syllable that

---

<sup>1</sup> Here and throughout this dissertation, I use the term "articulator" as shorthand for "primary oral articulator". The glottis, velum, and jaw are all also speech articulators (Brownman & Goldstein, 1990; Ladefoged, 1972), and two speech sounds that do not share a primary oral articulator can share other articulators, e.g., /v/ and /z/ have different primary oral articulators (lower lip and tongue-tip, respectively) but do have the same glottal articulation to produce voicing.

the subject was preparing to utter (the “identity” condition) than when the required response and distractor stimulus were different. When the response and distractor differed in this study, they differed only in articulator (e.g., *ba-da*).

### **2.3.3. Evidence from neuroimaging**

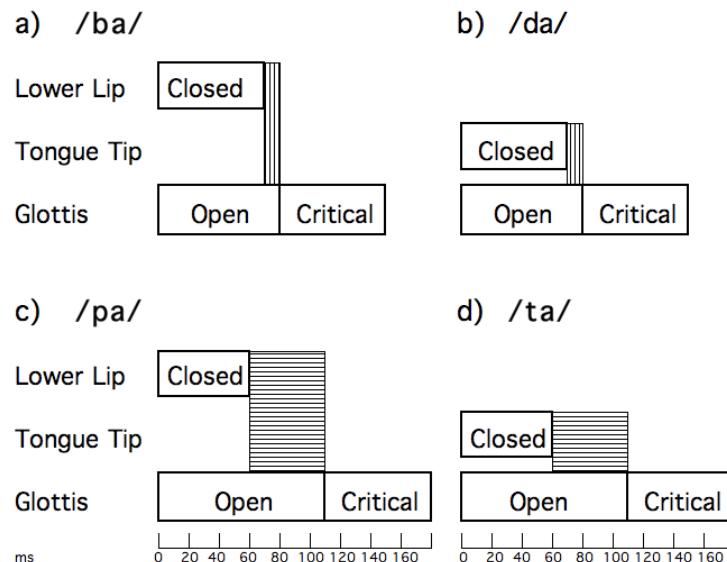
There is evidence from neuroimaging studies that brain areas that are associated with producing speech are involved in the process of perceiving speech. Pulvermüller et al. (2006) conducted a study using functional MRI and found that when subjects heard the phonemes /p/ and /t/, there was significant activation of motor regions in the precentral gyrus that were active when the same speakers moved their lips or tongue tip, respectively. The subjects did not show such increased activation in these areas when they heard other acoustic noise stimuli that were signal-correlated with the linguistic stimuli. The authors interpret this result as strong evidence that articulatory features of speech sounds are accessed during speech perception.

Gow Jr. and Segawa (2009) report an experiment that implicates speech motor areas of the brain (specifically, dorsal premotor area, dPM) as causing modulation of activity in areas of the brain known to be crucial for acoustic speech processing (specifically, bilaterally in posterior superior temporal gyrus, pSTG). Subjects were presented with audio of English two-word phrases where the final consonant of the first word of a phrase was underlyingly coronal (e.g., *pen tray*, /pən\_træɪ/). Such consonants are regularly assimilated to another place of articulation of a following

consonant, sometimes yielding real but different words (though somewhat nonsensical phrases), e.g., *gun pals*, [gumpælz], and sometimes yielding non-words (e.g., *pen pad*, [pɛmpæd]). Subjects had to indicate with a button press which of two pictures presented showed what they heard (nonsense words were represented by a red X). Gow Jr. and Segawa (2009) collected simultaneous EEG and MEG data from subjects with activity in identified regions of interest, with the localization of activity in each subject constrained by individual MRI scans of the participants (Dale et al., 2000). Their goal was to identify areas of the brain that caused activity in pSTG at around 200 ms. To do this they analyzed 40 Hz gamma phase locking patterns, which is believed “to reflect functional integration of ROIs across distributed cell assemblies” (Gow Jr. & Segawa, 2009, p. 225, and references immediately thereafter). In order to determine causality, they performed Granger analysis, a method that determines causal relations in signals across time (Granger, 1969; Kamiński, Ding, Truccolo, & Bressler, 2001), on these phase locking patterns. They found that when subjects heard nasals that had assimilated to the following non-coronal place of the following stop, activity in the dPM significantly influenced activity in the pSTG, but no such activity was noted when there was no assimilation (i.e., the following stop was also coronal). Gow Jr. and Segawa (2009) interpret this result as evidence that the speech motor control system is recruited in perception, and is especially strongly recruited to help disambiguate assimilated speech.

## 2.4. Discussion

In pointing to an intimate link between speech perception and production, the studies discussed above raise fundamental questions about how various properties of speech are involved in this link. One such question concerns the dimensionality and specificity of the representations involved in the perception-production link. For example, speech production involves spatial properties, i.e., which speech articulators make what constrictions where in the vocal tract, but also temporal properties, i.e., how those articulator movements are arranged relative to each other in time.



**Figure 1. Comparing temporal and spatial properties of a) ba, b) da, c) pa, d) ta.** The spatial property of interest is the primary oral articulator of the initial consonant, here made either with the lower lip or tongue tip. The temporal property of interest is the relative timing of the primary oral articulator and the glottis, represented by the boxes with bars. Rows (a and b, c and d) represent syllables with similar temporal properties but different spatial properties; columns (a and c, b and d) represent syllables with the same spatial properties but different temporal properties.

Consider the spatial and temporal organization of gestures for four CV syllables: *ba*, *pa*, *da*, and *ta*. Figure 1 shows representations of these four CV syllables using simplified gestural scores of Articulator Phonology (Browman & Goldstein, 1986, et seq.). From a spatial point of view, *ba* and *pa* are more similar than *ba* and *da* because *ba-pa* share a primary oral articulatory action (the closing of the upper and lower lips), but *ba-da* do not. In contrast, from the point of view of temporal organization, *ba* and *da* are more similar than *ba* and *pa*, because *ba-da* share a timing relation between the oral closure and the onset of glottal vibration, whereas *ba-pa* do not. This is witnessed by the similar VOT of around 0 ms for *ba* and *da*, while English *pa* typically has a VOT around 40 ms (Lisker & Abramson, 1964). This distinction between spatial and temporal properties of stimuli is highlighted because, surprisingly, past studies have found little experimental evidence for these properties of speech playing a role in the link between perception and production.<sup>2</sup> That is, whether temporal properties of speech play a crucial role in the perception-production link independently of spatial properties is unclear. One major goal of this dissertation is to address this question directly.

#### **2.4.1. Is identity special?**

The studies by Kerzel and Bekkering (2000) and Galantucci et al. (2009) found perceptuo-motor effects when perceived and produced utterances were the same, that

---

<sup>2</sup> The differences and similarities of these syllables could also be described as the phonological features of [±voice] and [labial]/[coronal] varying orthogonally.

is, when the perceived distractor syllables were segmentally identical to those to be uttered (referred to from here on as the “identity” condition). The Kerzel and Bekkering results suggest an effect of articulator, since the only cues available to the subjects from the video distractors were those of articulator. However, since subjects always responded *ba* or *da* and the silent video was always of a speaker saying *ba* or *da*, it is not possible to differentiate this result from that of Galantucci et al. where the response and distractor were auditory and identical. A fundamental question raised by these studies then is whether identity is unique in evoking perceptuo-motor effects, or whether shared temporal and spatial speech properties of voicing and articulator can independently yield perceptuo-motor effects. Several studies reviewed below have tried to find evidence, with little success.

Gordon and Meyer (1984) report on a series of experiments in which subjects learned sets of four cue-response pairs: a pair where the cue and response were identical (e.g., *pa-pa*), a pair where they shared voicing but differed in articulator (e.g., *ba-da*), a pair where they shared articulator but differed in voicing (e.g., *da-ta*), and a pair where they shared neither (e.g., *ta-ba*). Subjects heard the cue stimulus and spoke the learned response. RTs in the identity condition were significantly shorter than in the neither-shared condition. For the two conditions where the pair shared only one of either articulator or voicing the results differed: when the pair shared only voicing RTs were significantly shorter than in the neither-shared condition, but when the pair shared only articulator RTs were not significantly different from the neither-shared

condition. Gordon and Meyer (1984) concluded that speech perception shares a “common code” with speech production for voicing, but not for articulator. The lack of an effect of articulator is surprising given the undisputed status of both voicing and articulator in the description of linguistic contrasts (Chomsky & Halle, 1968; Ladefoged & Maddieson, 1996).

Galantucci et al. (2009) report another experiment (their experiment 1) using the same response-distractor task as in their experiment described in section 2.3.2. Articulator was either shared or different between the distractors and responses, but the distractors and responses always differed in voicing or nasality. They did not find shorter RTs when the distractors and responses shared articulator but differed in voicing (e.g., *ba-pa*) or nasality (e.g., *da-na*) than when they differed in articulator as well as in voicing or nasality (e.g., *ba-ta* or *da-ma*). That is, they found no effect for shared articulator. Again, the lack of an effect of articulator is surprising.

Mitterer and Ernestus (2008) attempted to elicit perceptuo-motor effects of articulator and voicing using a speeded shadowing task. In their experiment, subjects heard spoken Dutch stimuli consisting of two pseudowords of the form CVVC CVVC. Subjects were instructed to repeat both as quickly as they could. The onset consonant of the second item was of experimental interest. There were two manipulations. The first was based on a fact about Dutch /r/. The authors report that across Dutch speakers, this phoneme is realized as either the alveolar trill [r] or the uvular trill [R], but that a given Dutch speaker categorically produces only one or the other variant.

Half of the /r/-initial stimuli were produced with the alveolar trill and the other half with the uvular trill. All stimuli were produced by the same speaker (one of the authors, who they describe as special in her ability to produce both variants). Subjects were chosen so that one half were alveolar trillers and the other half were uvular trillers, though the authors experimentally verified that all subjects could discriminate between the two variants. The authors reasoned that perceptuo-motor effects of articulator should be obtained when subjects heard stimuli compatible to their habitual production, that is, RTs should be shorter for, e.g., alveolar trillers when they heard stimuli with alveolar trills but not when they heard stimuli with uvular trills. They found no such effect. RTs were not affected by stimulus type, response type, or an interaction between the two. The second manipulation concerned pre-voicing (i.e., negative VOT, where voicing starts during the closure of an obstruent). According to van Alphen and McQueen (2006), Dutch speakers are sensitive during lexical access to whether there is pre-voicing (i.e., negative VOT) in word-initial stops, but not to how much pre-voicing there is. Mitterer and Ernestus (2008) therefore created three levels of pre-voicing in stimuli starting with a voiced stop: 0, 6, and 12 cycles of glottal pulses before the release of oral closure. They reasoned that if the source of perceptuo-motor effects is activation of shared motor plans, then sub-phonemic differences in stimuli should be replicated or at least approximated by speakers in the shadowing task, as in Fowler et al. (2003). They did not find replication of the stimuli, rather, speakers produced significantly shorter VOTs in response to the stimuli with no

prevoicing compared with the responses to the stimuli with 6 or 12 cycles, but the latter two were not different from each other. Mitterer and Ernestus (2008) take these lack of results to signify that speech production is only marginally involved in speech perception, and that any link between the two systems is at the abstract, phonological level—not at the level of articulatory plans or gestures.

Facilitative effects other than those based on identity have also been elusive in speech-production experiments. Roelofs (1999) used a form-preparation paradigm to test whether the parameters of articulator or voicing facilitate spoken word production. His Dutch-speaking subjects learned three-word response sets, and a cue word that would indicate which response set to make on a given trial. Sets were divided into three block types. In identity blocks, the three responses always started with the same segment, e.g., *been* (“bean”), *bos* (“forest”), *baard* (“beard”). In baseline nothing-shared blocks, neither articulator nor voicing was shared across all three responses, e.g., *been*, *dolk* (“dagger”), *film* (“film”). There were two partial-matching block types, one in which one of the responses varied in voicing but always shared articulator with the other two, e.g., *been*, *bos*, *pet* (“cap”), and another in which one of the responses varied in articulator but always shared voicing with the other two, e.g., *vork* (“fork”), *been*, *damp* (“steam”). The dependent measure was the RT between presentation of the cue word and start of the first word of the response set. Roelofs (1999) found that only the identity blocks yielded facilitative effects on RTs in speech

planning compared to the baseline blocks. Partial-matching blocks based on voicing or articulator showed no facilitation compared to the baseline blocks.

In sum, there is little if any clear evidence in the literature for independent perceptuo-motor effects of articulator or voicing. It may therefore be the case that only segmental identity gives rise to perceptuo-motor effects. There is evidence that identity may have a special status in certain areas of phonology. Notably, the Obligatory Contour Principle has been argued (McCarthy, 1986; Yip, 1988, among many others) to prohibit co-occurring identical segments in both representations and derivational processes. Gallagher and Coon (2009) argue that segmental identity plays a crucial role in long-distance consonant harmony in Chol. Specifically, two co-occurring ejectives or two co-occurring non-ejective stridents in Chol roots are permissible only when they agree in all features, e.g., *sus* ‘scrape’ and *k’ok* ‘healthy’ are allowed while *\*ts-s* and *\*k’-p* are not. This is a requirement above and beyond feature-level co-occurrence restrictions in the language, e.g., that two stridents must agree in anteriority if one is ejective: *ts’is* ‘sew’ is allowed but *\*ts’-f* is not, which they attribute to articulatory locality (Gafos, 1999). Though the Obligatory Contour Principle and the long-distance harmony differ in whether they penalize or require identical segments, they both crucially refer to the notion of segmental identity. It may be the case, then, that only representations at the level of segmental identity are involved in the perception-production link.

On the other hand, there are reasons to question the conclusion that only the identity condition should have a role in the link between speech perception and production. Phonological patterns in the world's languages exhibit systematicities in terms of phonological features, which include the parameters of voicing and articulator. Many languages restrict which consonants can co-occur in a morpheme based on the laryngeal features of the consonants, e.g., voicing (see MacEachern, 1999 for an extensive review). For example, a certain class of native Japanese morphemes cannot contain more than one voiced obstruent consonant, that is, morphemes like *\*gaze* are not allowed (Itô & Mester, 1986). Patterns based on articulator are also well-attested. Many languages show an aversion to morphemes containing consonants that share articulator, most notably in Arabic (Frisch & Zawaydeh, 2001; Greenberg, 1950; McCarthy, 1986), and also in a wide variety of other languages (see Frisch, Pierrehumbert, & Broe, 2004, for a review). Phonological patterns based on voicing are widespread. For example, clusters of obstruent consonants in Russian agree in voicing with the voicing of the last consonant in the cluster (see Halle, 1971, among many others), e.g., *pojezda* vs. *pojest* 'train (genitive vs. nominative singular)'. Given these widespread phonological effects based on voicing and articulator, it is reasonable to expect perceptuo-motor effects of voicing or articulator, especially since there is a long line of work (e.g., Hura, Lindblom, & Diehl, 1992; Ohala, 1993, 2005) arguing that sound patterns of languages have their bases in the relation between perception and production. Since the advent of generative phonology (Chomsky &

Halle, 1968; see also Kenstowicz, 1994) phonological theory has relied heavily on features rather than segments in accounting for the sound patterns of languages. It would therefore be surprising if the properties of speech that play such a critical role in phonological theory played no role at all in the link between speech perception and production.

There is reason to suspect that issues of experimental design may be responsible for the lack effects of articulator and/or voicing in the studies reported above. In the experimental designs used by Gordon and Meyer (1984) and Galantucci et al. (2009), subjects had to decide on a given trial which response to make based on some stimulus. On any given trial the speaker's selection of a response could have been biased by hearing a distractor somehow similar to one of the potential responses before he or she had actually selected which response to make. Such stimulus-response compatibility (Kornblum, 1994) can result in quicker selection of the required response and therefore faster RTs. The key is that, in such stimulus-response compatibility effects, the similarity leading to faster selection might not be limited to simply sharing an articulator. It could be acoustic, visual, or some other type (or combination of types) of similarity. These "selection effects" (Galantucci et al., 2009; Kerzel & Bekkering, 2000) do not necessarily implicate the involvement of the speech production system during the perception of the distractor. To ensure that differences in RTs reflect perceptuo-motor effects rather than selection effects, the distractor can be presented after the cue but before the response syllable prompted by the cue is

completely planned and executed. By delaying the onset of the distractor by a couple hundred milliseconds after the cue is presented (i.e., by employing positive SOAs), any influence of the distractor on the speakers' response selection can be precluded. Any distractor influence is then due to mechanisms common to or mediating between perception and production. The results obtained by Kerzel and Bekkering (2000) and Galantucci et al. (2009) for the identity condition were significant at positive SOAs, and therefore likely reflect perceptuo-motor effects.

However, in the experiment where Galantucci et al. (2009) manipulated voicing/nasality between distractors and responses and did not find a result, negative SOAs were used. Their experimental design therefore may not have been appropriate for eliciting perceptuo-motor effects. It is possible that the result from Gordon and Meyer (1984), where shorter RTs were found when the cue and response shared voicing but differed in articulator, also reflected selection effects since by their learned cue-response pair design, the cue indicated which response the subject was to select on a given trial. The experimental task used by Mitterer and Ernestus (2008) is also ill-suited for eliciting perceptuo-motor effects, since the task they used by its nature must include the influence of selection processes on RTs. Subjects did not know what word they were going to produce until they heard it spoken. Whether or how features influence the selection process is not the question being addressed in this dissertation. In the task used by Roelofs (1999), demands of buffering the queue of responses in memory as well as requirements of speech production were involved in the task since

the required responses triplets had to be memorized in advance of the experimental trial blocks. It is unclear from Roelofs's results what the relative influence of these memory demands vs. that of speech production were, so the lack of evidence in favor of a role for features should not be taken as obviously problematic for the view that perceptuo-motor effects of articulator and/or voicing should be obtainable.

#### **2.4.2. Speech perception debate**

A longstanding debate exists as to whether the speech production system is involved in the process of speech perception (see, e.g., Diehl et al., 2004, for a review). One class of theories of speech perception (Fowler, 1986; Galantucci et al., 2006; Liberman & Mattingly, 1985) asserts that the articulatory gestures used to produce speech are obligatorily involved in the process of speech perception. Other authors (e.g., Hickok, 2008; Lotto, Hickok, & Holt, 2009; Mitterer & Ernestus, 2008; Ohala, 1996) claim that speech perception need not necessarily rely on the speech production system. The most vigorous objections of the latter authors with the theories of the former are to the claimed obligatory nature of the involvement of the production system in perception. Whether such involvement is necessary is an independent question from the perception-production link evidenced by the studies, a matter on which there is little or no disagreement in the literature. Assuming that the shorter RTs in the studies described in 2.3.2 and 2.3.2 were due to increased activation of motor plans due to perceived stimuli, the question of whether the increased activation was

due to the obligatory recruitment of the speech production system in perception or due to indirect associations between separate representations involved in perception and the motor system (cf. Viviani, 2002) or to separate parallel processes for auditory-lexical and auditory-motor mapping (Hickok & Poeppel, 2000, 2007), the link seems to be real and appears to be involuntary.

## 2.5. Conclusion

There is ample evidence in the literature for an interaction between speech production and speech perception, but the experimental evidence is inconclusive as to the dimensionality and specificity of the representations that are involved in that link. One crucial goal of this dissertation then is to examine whether perceptuo-motor effects of voicing and articulator can be found. It is necessary to determine whether such effects exist in order to specify the minimal requirements that any model of the perception-production link must satisfy. Results from two experiments are presented in the next chapter. The experiments were inspired from previous studies on perceptuo-motor interaction, but expanded on these by dealing explicitly with the methodological issues discussed above. Thus, the experimental task was the same response-distractor task used by Galantucci et al. (2009), but with stimuli designed to manipulate the temporal and spatial parameters of voicing and articulator independently. In addition, the experiments made use of positive SOAs only with the aim of exposing perceptuo-motor effects.

The results of the experiments presented in the next chapter then serve to inform the development of a formal computational model of the process of phonological planning. This model will account for the results of previous studies and the results from the next chapter by explicitly focusing on the dynamics involved in phonological planning.

## CHAPTER 3: EXPERIMENTS

### 3.1. Introduction

This chapter presents two experiments that sought evidence of perceptuo-motor effects of voicing (experiment 1) and articulator (experiment 2). As discussed in Chapter 2, previous studies that have found experimental evidence for perceptuo-motor effects have done so when a perceived utterance was identical to an utterance being planned (Galantucci, Fowler, & Goldstein, 2009; Gordon & Meyer, 1984; Kerzel & Bekkering, 2000). Gordon and Meyer (1984), Mitterer and Ernestus (2008), and Galantucci et al. (2009) all ran experiments attempting to uncover perceptuo-motor effects of articulator, all unsuccessfully. Only Gordon and Meyer (1984) sought experimental evidence for perceptuo-motor effects of voicing, and did find some evidence for them. However, as discussed in section 2.4.1 of Chapter 2, there are some considerations of experimental design that make interpreting the Gordon and Meyer (1984) result conclusively as representing perceptuo-motor effects questionable.

Experimental evidence from speech production also points to a special role for identity. Roelofs (1999) used a form-preparation paradigm to test whether the properties of articulator or voicing facilitated spoken word production. Roelofs (1999) found that only full segment identity yielded facilitative effects in speech planning. Segments differing only in voicing or articulator did not provide any such facilitation. If speech production is insensitive to the properties of articulator and voicing, then it

would not be surprising for any link between perception and production to likewise be insensitive to these properties.

It may be the case that only the identity condition can give rise to perceptuo-motor effects. Alternatively, the lack of evidence for perceptuo-motor effects of articulator/voicing in previous studies might have been due to not testing any cases other than identity (e.g., Kerzel & Bekkering, 2000), or not testing optimally for perceptuo-motor effects (Galantucci et al., 2009; Gordon & Meyer, 1984). Results from two experiments that addressed this question directly are presented below. The experimental task was the same response-distractor task used by Galantucci et al. (2009). The experiments reported here carefully manipulated the articulator and voicing of the stimuli and responses independently while excluding the identity condition. Only positive SOAs were used to avoid selection effects (see section 2.4.1 of Chapter 2 for details). The first experiment tested for perceptuo-motor effects of voicing, and the second for effects of articulator. The hypothesis to be tested in each was that RTs on trials where the response and distractor share a particular property should be shorter than on trials where they differ on that property. It is necessary to determine whether such effects exist in order to specify the minimal requirements that any model of the perception-production link must satisfy.

### **3.2. Voicing RT experiment (Experiment 1)**

This experiment tested whether a perceptuo-motor effect of voicing could be identified. If such an effect is present, then RTs on trials when the response and distractor match in the temporal property of voicing should be shorter than trials when they differ in voicing.

#### **3.2.1. Methods**

All materials and procedures were approved by the Institutional Review Board of New York University (the University Committee on Activities Involving Human Subjects).

##### **3.2.1.1. Subjects**

49 subjects were recruited from the subject pool of the New York University Department of Psychology, and received credit in fulfillment of class requirements. All subjects identified themselves as native speakers of American English, and as having no speech or hearing impairments. Eight speakers had been exposed to another language in the home from early childhood. A technical issue resulted in no audio recording for one subject. One subject had a head cold and yawned so frequently in the recording session that an extremely large number of trials were not analyzable. One subject consistently pre-nasalized a very large portion of responses. Data from these 11 subjects were therefore excluded from analysis. Data from the remaining 38

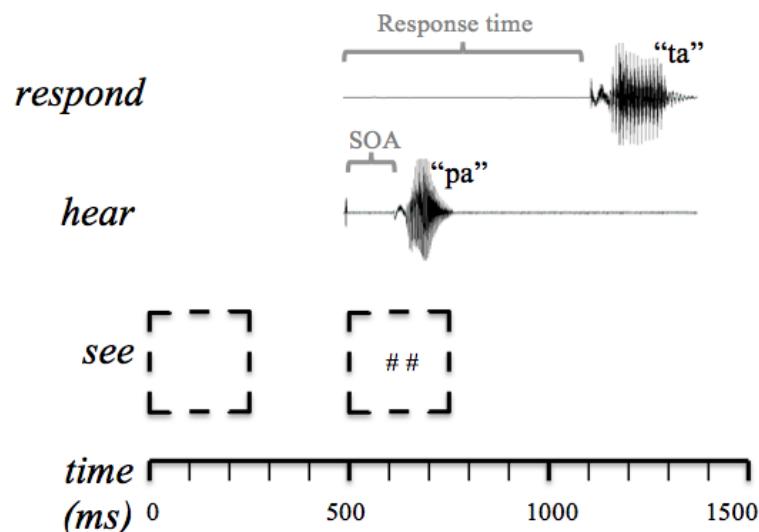
subjects (12 male, 26 female) were analyzed. All subjects gave informed consent before participating in the experiment.

### **3.2.1.2. Procedure**

Subjects sat in front of a computer monitor in a sound-attenuated booth in the Phonetics and Experimental Phonology Lab at New York University while wearing headphones. Subjects were told that they would learn sets of symbol-syllable pairs, and that in the experiment the computer would repeatedly show one of the two symbols. They were told that their task was to say the syllable that they had learned for that symbol. They were also told that while they were performing that task, they would hear various sounds over the headphones, which they should ignore. They were told to respond as quickly as possible, but not so quickly that they made a lot of mistakes. They were told that there were two microphones, one to record their response and another to register that they had responded on each trial. They were told that the computer would not proceed to the next trial if it did not register that they had responded to the current trial. Therefore, they should speak louder into the second microphone if the computer did not move to the next trial.

The subjects had two practice blocks before beginning the experiment. The first consisted only of a sample instruction screen that showed them the symbol-syllable pairing for that block, followed by eight trials without audio distractors. This block served to introduce the basic task to the subject, and to ensure that the

microphone recording that the subject had responded on each trial was at an appropriate distance from the subject to reliably detect their replies. The second practice block introduced the audio distractors. After completing the second practice block, subjects were told that they would then be presented with three blocks, each lasting about ten minutes. Each block started with an instruction screen giving them the symbol-syllable pairing for that block, which would be different for each block. They were told that they could rest when they saw the instruction screen for the next trial. The blocks for this experiment were the first and the third. The second block was for a different experiment.



**Figure 2. Timeline of one trial. A fixation box was presented for 500ms, at which point the visual cue for the trial was presented, here # # indicating a response of ta. A short audio burst was played simultaneously with the presentation of the visual cue. An audio distractor was played with one of three SOAs, here 100 ms.**

The timeline of a trial with a congruent distractor is illustrated in Figure 2. In each block, the two possible responses were stop-vowel syllables where the stops matched in articulator but differed in voicing (e.g., *ta-da*). Each trial consisted of a fixation box that appeared on screen empty for 500 ms. The visual symbol cue (either == or # #) would then appear in the box. After the presentation of the visual cue, some trials had an audio distractor and some did not. The response-distractor pairs are shown in Table I. Any general effect of voicing should be present across multiple articulators. Therefore, responses with two different articulators were included in the design, a block with tongue-tip responses (*ta-da*) and a block with tongue-back responses (*ka-ga*). 15 of the subjects saw == as the symbol cue for the voiced response, 23 saw == as the symbol cue for the voiceless response. The imbalance was

**Table I. Response-distractor pairs for the voicing experiment.** A green check mark (✓) indicates congruent voicing, a red X incongruent voicing. Dots show the tone and no distractor conditions, which were also included for each block. In the first experimental block, subjects responded either with *ta* or *da* (tongue-tip block, white background). In the third experimental block, subjects responded either with either *ka* or *ga* (tongue-back block, grey background). All distractors were presented at SOAs of 100, 200, and 300 ms. Trials without distractors could not manipulate SOA, as there was no second stimulus.

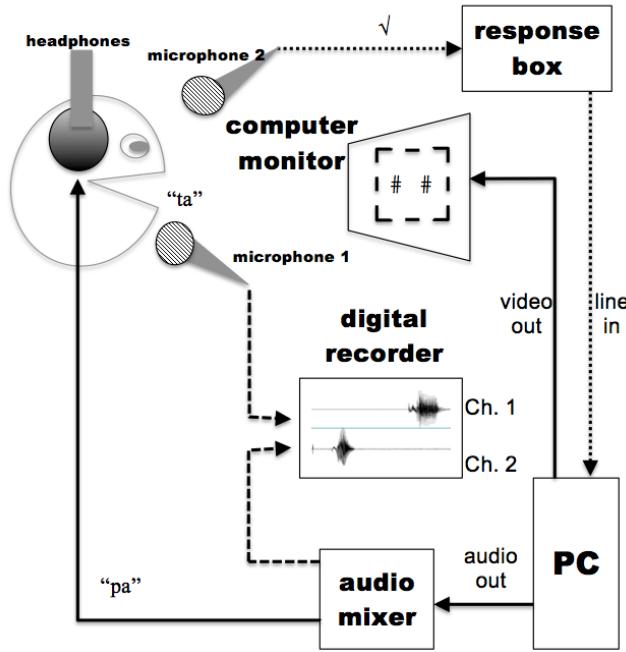
| Response                 |           | Distractor |           |             |             |
|--------------------------|-----------|------------|-----------|-------------|-------------|
|                          |           | <i>pa</i>  | <i>ba</i> | <i>tone</i> | <i>none</i> |
| Block 1<br>(tongue tip)  | <i>ta</i> | ✓          | X         | •           | •           |
|                          | <i>da</i> | X          | ✓         | •           | •           |
| Block 2<br>(tongue back) | <i>ka</i> | ✓          | X         | •           | •           |
|                          | <i>ga</i> | X          | ✓         | •           | •           |

due to more of the excluded subjects being in the ==voiced group. Simultaneous with the onset of the visual cue specifying the required response to the subject, a very short (15ms) high-frequency tone was played. This sounded like a very quiet, short burst of static over the headphones. This tone was used for the hand-measurement of RTs.

The audio distractors were presented with an SOA of 100, 200, or 300 ms after the onset of the visual cue. Each response/distractor combination was presented 14 times at each of the three SOAs, yielding 252 trials per block. There were 28 trials in each block (14 per response) where the subjects did not hear a distractor, bringing the total to 280 trials per block, 560 trials per experiment. The computer kept the cue on screen until a response was recorded from the subject on that trial, as detected by the hardware voice key in the serial response box attached to the second microphone. 1000 ms after the detection of a response on a given trial, the next trial began. Trials were pseudo-randomized within block. The entire experiment including the two practice blocks and three experimental blocks lasted about 40 minutes per subject.

### **3.2.1.3. Equipment**

A schematic drawing of the experimental setup is shown in Figure 3. The visual cue stimuli and audio distractor stimuli were presented to subjects from a PC using e-Prime Professional 2.0 (Psychology Software Tools, Inc., Pittsburgh, PA). Audio output from the PC was fed into an Oz Audio Q-mix HM-6 headphone matrix amp audio mixer, which in turn sent the audio output both to a set of Sennheiser



**Figure 3. Schematic of the hardware setup for the experiments. Software running on a PC controlled presentation of the visual cues and audio distractors to the subject (solid arrows). Audio output from the software and verbal responses from the subject were recorded by a digital audio recorder (dashed arrows). A hardware voice key in the response box detected when a response had been made and indicated to the software that the next trial should begin (dotted arrow).**

HD201 binaural headphones worn by the subject and to one channel of a digital Zoom Handy H4n Recorder. The other channel of the digital recorder was attached to an Audio-technica AT2010 microphone that recorded the entire experimental session at 44.1 Hz and 16-bit sampling rate, synchronously with the audio output from the experiment (the same as what the subject heard). A second microphone (Audio-technica ATR20) was attached to a serial response box that detected when the subject had responded on each trial, by means of the hardware voice key provided by e-Prime. The RTs recorded by the voice key were not used in data analysis (see section 3.2.2),

but rather to determine that the subject had replied and that the next trial should be presented.

### 3.2.1.4. Stimuli

Distractor syllable stimuli were recorded by a female native speaker of American English, who was naïve to the purpose of the experiment. The stimuli were recorded in the sound-attenuated booth of the Phonetics and Experimental Phonology Lab at New York University using a digital Zoom Handy H4n Recorder at 44.1 Hz and 16-bit sampling rate. The syllables were spoken in isolation by the speaker as they were presented in English orthography randomly on a computer screen. Each syllable was produced five times. The tokens that were most conducive to creating consistent stimuli based on the properties discussed below were chosen. Stimulus files were then created by the following procedure.

**Table II. Properties of the distractor stimuli in the voicing experiment. VOT was the time between the beginning of the stimulus file and the onset of periodic energy corresponding to the phonation associated with the vowel. Vowel Duration was the section of the stimulus file from the onset of that periodic energy to the end of the stimulus file.**

| stimulus  | vowel | format   | F0 (Hz)              | F1 (Hz) | F2 (Hz)  |
|-----------|-------|----------|----------------------|---------|----------|
|           | VOT   | duration | transitions          | range   | end      |
| <i>ba</i> | 0 ms  | 150 ms   | ~27 ms<br>(5 pulses) | 179-193 | 858 1431 |
| <i>pa</i> | 40 ms | 111 ms   | ~17 ms<br>(4 pulses) | 167-205 | 914 1352 |

The sound files were edited so that they would be 150-151 ms in duration, making sure that the beginning and end of the excised section was at zero amplitude crossings in the waveform. For the voiceless stop *pa*, the beginning was chosen so that the stimulus had a VOT of 40 ms. For the voiced stop *ba*, the beginning of the stimulus was chosen as close to the release of the stop as possible, so that the stimulus had a VOT of 0 ms. The intensity at the end of each sound file was attenuated to make the stimulus sound more natural in isolation. The intensity of the final 65 ms of each file was modified to decrease from the full intensity at 65 ms before the end of the file to 5/6 of the intensity at the end of the file. Sound files were then modified to have a consistent average intensity. The stimuli used in the present experiment were played to native speakers of American English, who easily identified the syllables correctly.

Table II shows the durations and other properties of the distractor stimuli files.

Each stimulus file started at the release of the oral closure marked by the release burst, and each file had a vowel whose duration was the amount of time from the beginning of periodic energy associated with phonation to the end of the stimulus. The distractor stimuli (*ba/pa*) differed in the voicing of the initial consonant because the purpose of the experiment was to test for effects of (mis)match in voicing between the distractor and the response. This meant that the *pa* distractor had to include acoustic material corresponding to the 40ms VOT that the *ba* distractor did not. Therefore, a choice had to be made between either keeping the vowel duration

constant or overall stimulus duration constant across stimuli. Two considerations strongly favored keeping the overall stimulus duration constant. First, the results of Galantucci, Fowler, and Goldstein (2009) suggest that longer distractors result in longer response latencies in this task. In their experiment 1, their distractor stimuli were approximately 400 ms in duration, while in their experiment 2, they were 150 ms. Response latencies in their experiment 1 were approximately 550 ms, while in experiment 2 they were approximately 470 ms.<sup>1</sup> Keeping the distractors the same length should avoid introducing this effect. Second, since the acoustic cues relevant to the experimental manipulation (VOT) are at the beginning of the stimulus, the distractor length was held constant and the vowel length was shorter for the *pa* distractor than the *ba* distractor. Since the time between stimuli (SOA) was also tightly controlled and manipulated in the experiment, it was viewed more desirable to have as many other temporal factors held constant as possible from trial to trial.

### 3.2.2. Results

38 subjects yielded 21280 trials. Data were excluded based on certain criteria. Occasionally e-Prime introduced delays in the presentation of some stimuli that resulted in the actual SOA differing materially from the intended SOA. A trial was

---

<sup>1</sup> The effect of distractor length was likely larger than it seemed in their data. The difference in response latencies was probably mitigated by other differences between their two experiments: the SOAs for their experiment 1 were negative (the distractor was presented before the visual cue), while in their experiment 2 they were all positive (the distractor was presented after the visual cue). Given the robust effect of SOA on response latencies that they found, whereby larger SOA reliably increased response latency, the difference in RTs would be expected to be even larger had both experiments been run at comparable, longer SOAs.

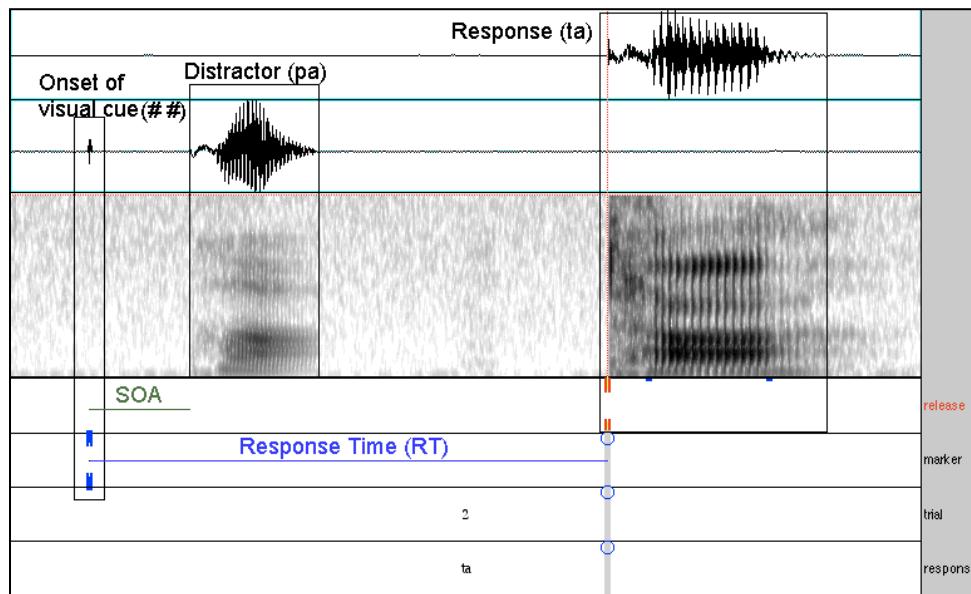
excluded from analyses if its actual SOA differed by more than 30 ms from the specified SOA. A trial was also excluded if the subject did not produce the correct syllable, based on the judgment of the experimenter. The experimenter was unaware of what the correct response on a given trial was supposed to be. Responses were marked as incorrect only when the subject said something other than one of the syllables in that block. For example, in the block with the possible responses of *ka* or *ga*, responses of, e.g., *gaka*, *ba*, a yawn, etc. were marked as incorrect. However, if in that same block the correct reply was *ga* but it impressionistically sounded like *ka* to the experimenter, it was not excluded as erroneous since variations in VOT were expected in the data. In other words, tokens were not excluded based on variations in VOT that might have seemed incorrect to the experimenter but were intended productions of *ka* but with shorter than normal VOT.

Since the intent of the experiment was to test for perceptuo-motor effects during production, certain responses were excluded from analysis based on RT. Trials were excluded if the subject's response started earlier than 100 ms into the playing of the audio distractor. It does not seem reasonable to expect the distractor to evoke perceptuo-motor effects if the subject started a production before he or she had time to perceive the audio distractor. It is a common practice in response-time experiments to exclude responses where the subject takes too long to reply. This is often done by excluding trials that are longer than 2.5 or 3 standard deviations of the within-subject mean. Baayen (2008, p. 244) notes that it is preferable not to exclude data on that

basis, as this criterion runs the risk of removing legitimate observations from a normal distribution. Baayen (2008) recommends instead picking some other principled criterion. In this analysis, trials were excluded from analysis if the RT was greater than 750 ms after the presentation of the distractor, on the assumption that the subject was inattentive on that trial. The average RT across all remaining trials was 548 ms.

Lastly, it is nonsensical to refer to an SOA on trials where subjects heard no distractor, since there was only one stimulus on those trials. Therefore they were assigned an SOA of “N/A”. Since both SOA and Distractor were included as fixed factors in the statistical analysis, the factors should not be correlated. The correlation between SOA of “N/A” and the blank Distractor is 1. The blank distractor can provide insight into the effects of the presence vs. absence of a distractor, but is not necessary for comparing the effects of the congruency of a distractor-response pair. The blank trials were therefore also excluded from the statistical analyses below. In total, 5053 trials (23.75%) were excluded based on these criteria, leaving 16337 trials to be included in the analyses.

RTs were measured by hand using Praat (Boersma & Weenink, 2006), rather than using the hardware voice key provided by e-Prime, as Rastle and Davis (2002) found hardware voice keys based solely on a pre-set intensity threshold to be systematically unreliable. Figure 4 shows the landmarks used to measure RTs available in the stereo recording.

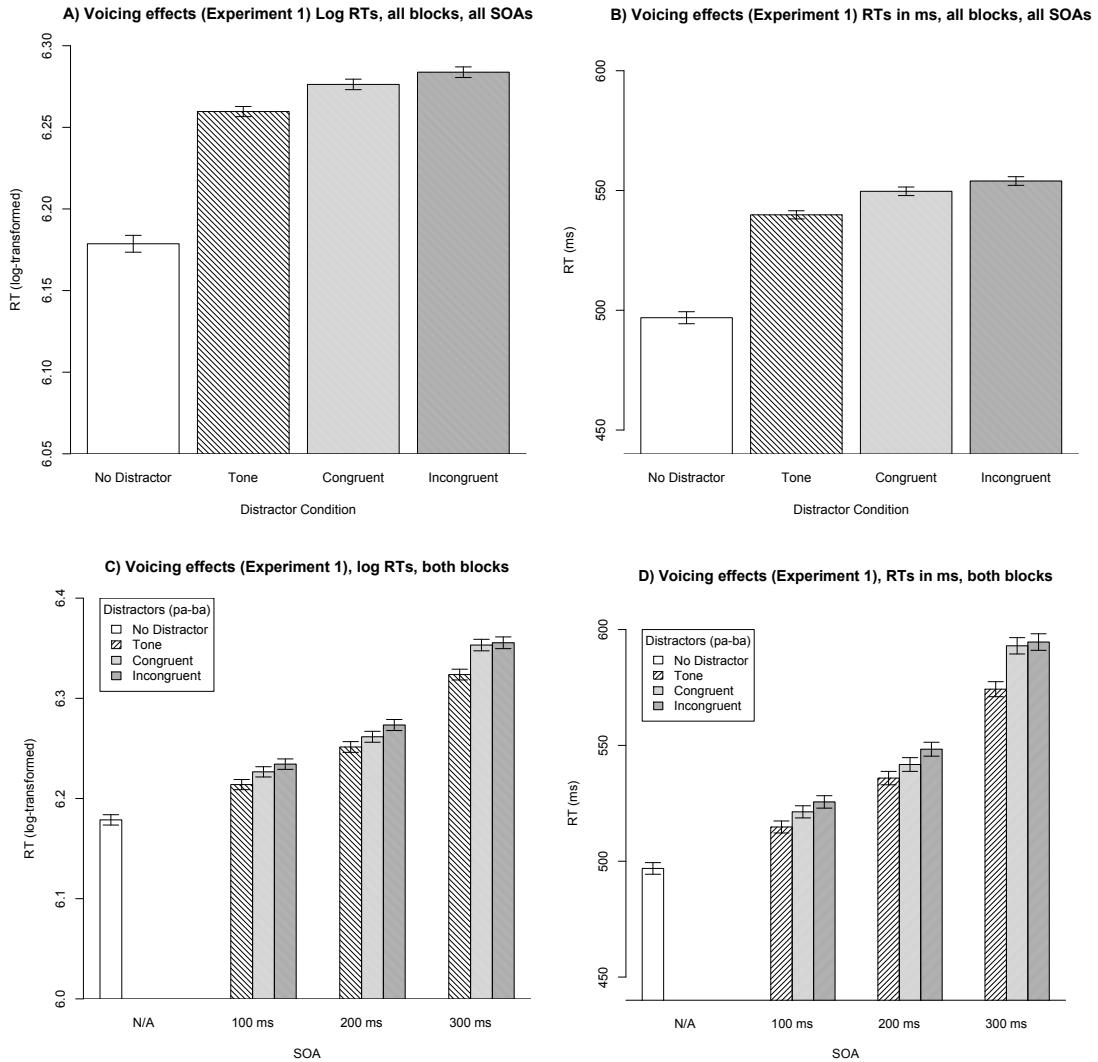


**Figure 4. An example of the data labeling for one token of a subject replying *ta*.** The leftmost vertical box shows the acoustic record of a short tone that was synchronous with the onset of the presentation of the visual cue that told the subject which syllable to say. The second vertical box shows the acoustic record of the distractor. The third vertical box shows the subject's response on the other recording channel. The onset of the response was coded as the beginning of aperiodic energy associated with the release of the stop /t/. Response time was calculated as the time point of response stop release minus the onset of the visual cue.

Ideally, RT would be measured as the time between the presentation of the visual cue and the onset of articulatory movement associated with syllable. The collection and analysis of dynamic articulatory data, e.g., using electromagnetic articulography or ultrasound, was unfortunately time- and cost-prohibitive given the amount of data that needed to be collected for the present experiments. RT was measured as the latency between the onset of the visual symbol cue (# # or ==) as recorded by the simultaneous tone (adjusted by the tone presentation delay as recorded by e-Prime) and the release of the stop closure, as discernible in the waveform of the

subject's response. The use of the release of the oral closure was chosen as the acoustic landmark for all trials because it was common to all trials, and because Mooshammer et al. (2012) found a systematic and reliable relation between the onset of the acoustic signal and the onset of articulatory movement for stop-initial CV syllables.

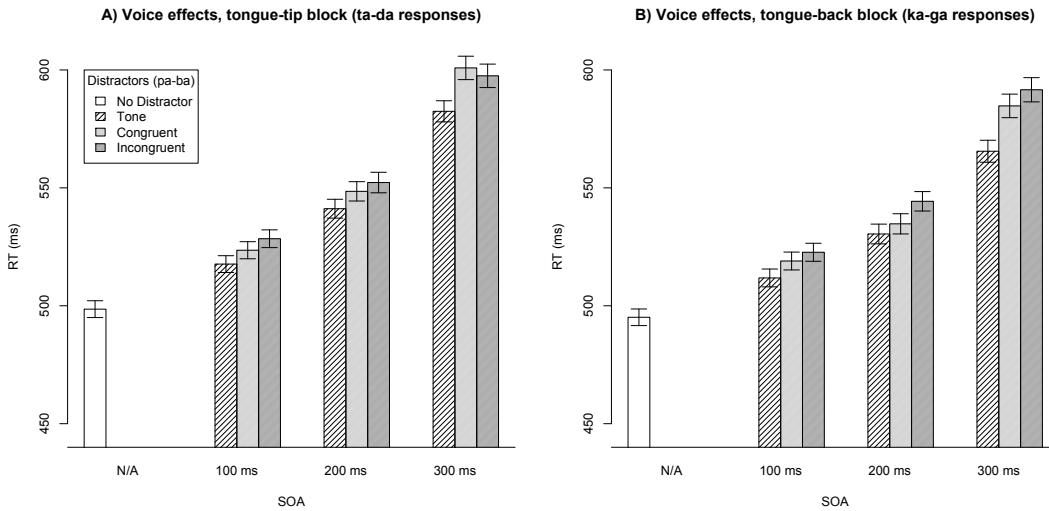
The hypothesis to be tested with these data was that RTs on trials where the response and Distractor shared voicing (Congruent) should be shorter than those where they differed in voicing (Incongruent). Figure 5 shows the mean RTs for each Distractor condition, broken down by SOA. Means in Figure 5A are shown for the log-transformed RTs across all SOAs, which were subject to statistical analyses.



**Figure 5. Mean RTs for the voicing experiment (Experiment 2). The linguistic Distractors were *pa* or *ba* in both blocks.** A) shows the means of the log-transformed RTs by Distractor across both blocks and all SOAs, which were subject to statistical analyses. B) shows the mean RTs in milliseconds across all SOAs for ease of interpretation. Error bars show 95% confidence intervals. The No Distractor condition was not included in the statistical analyses, but is shown here for comparison. C) shows the log-transformed RTs from A) by SOA. D) shows the ms RTs from B) by SOA.

Figure 5B shows the means in milliseconds in order to give a more intuitive sense of the RT differences, again across all SOAs. Figure 5A shows that the

hypothesis that RTs in the Congruent condition should be shorter than in the Incongruent condition was supported. The statistical analyses below show that this difference was significant.



**Figure 6. Mean RTs in ms for Experiment 1, by block. A) shows the RTs for the block with tongue-tip responses, and B) shows the RTs for the block with tongue-back responses. Distractors were always *pa* or *ba*.**

Figure 6 shows that the general pattern of RTs is roughly the same within each of the two experiment blocks, with the single exception of the tongue-tip block at 300-ms SOA.

### 3.2.2.1. Statistics

Interpreting response-time data is challenging in that many factors contribute to how long it takes a subject to reply on a given trial, even in a relatively simple design like that of the present experiment (Baayen, 2008). The statistical tools used to analyze the data from the present experiment should ideally have two features. First,

they must be able incorporate the multiple factors that influence RTs, both those introduced by experimental manipulation as well as those that are not. Second, they should be compatible with the practical realities of collecting this type of data. Linear mixed-effects modeling, referred to simply here as "mixed models"<sup>2</sup>, (Baayen, Davidson, & Bates, 2008; Barr, Levy, Scheepers, & Tily, under review; Max & Onghena, 1999) meets both of these criteria. A brief explanation of why other common statistical analyses were either inappropriate or insufficient is followed by an overview of the benefits of mixed models. Details of mixed modeling are presented in the next section in the context of modeling the present results.

Simple univariate ANOVA was inappropriate for analyzing the present data because it assumes that each subject contributes only one observation to the data. In the present experiment, this is not the case: each subject contributes hundreds of observations to the data. That is, there are repeated measures taken from each subject. Assuming that one takes care to correct the degrees of freedom used in the analysis (for discussion see Max & Onghena, 1999), ANOVA with Repeated Measures (RM-ANOVA) can appropriately be used. However, in RM-ANOVA, it is assumed that subjects contribute one observation to every cell of the experimental design. If, as in the present data, subjects contribute more than one observation to each cell, RM-ANOVA requires that the subject's responses be averaged within-cell, effectively being treated as one observation. This approach has two drawbacks. The first is that

---

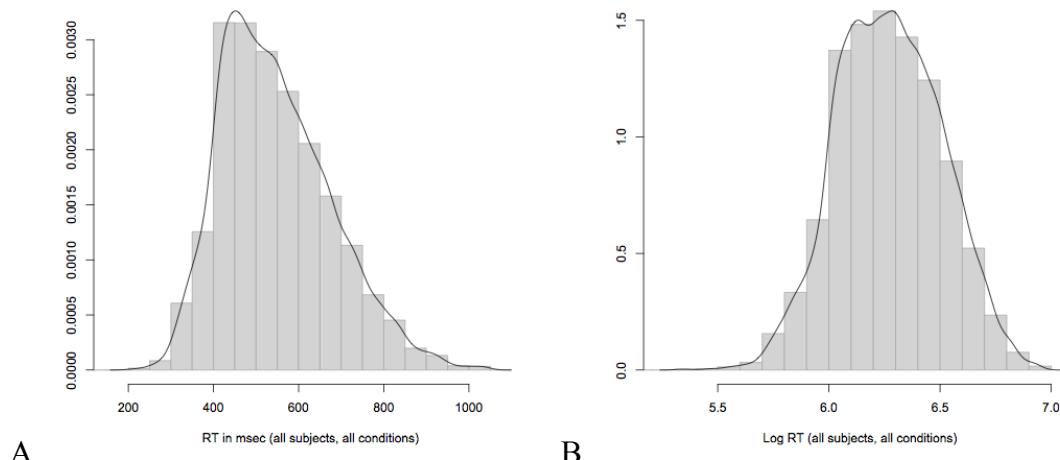
<sup>2</sup> Gelman and Hill (2007) object to this term in preference to "hierarchical/multilevel models".

the within-cell subject variance is lost in the analysis. The second is that certain task effects that are known to affect RTs are also lost. For example, let us make the assumption that all subjects get tired as the experiment progresses (which is not straightforwardly the case, as will be shown in section 3.2.2.2.1). Since the presentation of stimuli is pseudo-randomized, if it turned out across subjects that Incongruent trials on average appear later in the experiment than Congruent trials, we could incorrectly decide that it was congruency that resulted in slower responses, when in fact it was simply the subjects getting tired. It is preferable to have a way to factor these effects into the analysis, which is impossible in this design with RM-ANOVA.

Mixed-effects models are so called because they allow any number of random effects and fixed effects to be included in the same statistical model (see Baayen, 2008; Baayen et al., 2008; Barr et al., under review; Bates, 2005; Johnson, 2008, for discussion). Repeated measurements from the same subject can be used, avoiding the problems with ANOVAs raised above. Mixed-effects models also allow for the inclusion of longitudinal effects like those discussed above. Both continuous (e.g., RT of the previous trial) and categorical effects (Distractor condition) can be included in the same model. Mixed models do not assume balanced datasets, which is important for these data given the exclusion criteria outlined above. All of these features make mixed models a powerful and particularly appropriate tool for analyzing response-time data. Mixed models use fixed and random effects to account for variance in the data.

For each fixed effect, a slope coefficient is calculated, which indicates the magnitude and direction of influence that the effect has on the dependent variable. An  $R^2$  can be calculated for the overall model, which indicates how much of the variance in the data can be accounted for given the random and fixed effects included in the model. A model that accounts for more of the variance is a better model, assuming that the fixed effects included in each are significant.

It is an assumption of mixed models (as it is in RM-ANOVA) that the dependent variable values are normally distributed.<sup>3</sup> Response-time data in general are not normally distributed, but positively skewed and bounded on the negative side, because non-erroneous response-time data cannot be less than 0. This can be seen in the present data in Figure 7A.



**Figure 7. A)** The non-transformed response-time data for all subjects and conditions was positive skewed. **B)** The same data log-transformed was closer to a normal distribution.

---

<sup>3</sup> though (Gelman & Hill, 2007) maintain that this is not an important assumption.

Log-transformed data (Figure 7B) of the present data was closer to a normal distribution than the non-transformed data. Table III show the results from Kolmogorov-Smirnov tests of the two distributions. The very low  $p$  values show that neither distribution was normal. However, comparing the  $D$  values of the distributions, which indicate distance from a normal distribution, the  $D$  value of the log-transformed data was about half that of the non-transformed data. The log-transformed data was therefore closer to a normal distribution, and more appropriate for the planned statistical tests. References to “RT” and “Response Time” in the statistical models and analyses from here on refer to the log-transformed RTs, unless explicitly noted otherwise.

**Table III. Results of Kolmogorov-Smirnov tests for normality on the non-transformed RTs in milliseconds and log-transformed response-time data.**

|         | $D$   | $p$      |
|---------|-------|----------|
| RT msec | 0.057 | < 0.0001 |
| Log RT  | 0.032 | < 0.0001 |

An advantage of mixed models is that they can include any number of “control effects”—that is, effects that do have an influence on RTs, but are not necessarily of any particular theoretical interest—as well as “experimental effects”, which are of theoretical interest. Barr et al. (under review) point out that including these control effects in a model “can rule out potential confounds and increase statistical power by reducing residual noise” (Barr et al., under review: p. 60). Some authors (Baayen,

2008; Baayen et al., 2008) advocate building mixed models incrementally, that is, by starting with a simple model including only a small number of random and fixed effects, and adding only those effects that significantly improve the amount of variance accounted for by the model with more factors. However, Barr et al. (under review) show that it is preferable to use a model with maximal number of random effects, as it is not appropriate to use an ANOVA model comparison to determine whether to include a given random slope. In particular, Barr et al. (under review) recommend that the mixed-effect models be structured such that subject- (and item-) specific terms be included to accommodate variation of fixed effects across random effects. In other words, the model should ideally be constructed so that it factors in the possibility that not all subjects will behave uniformly with regard to the fixed effects (control and experimental). Inclusion of all of these random effects can often lead to models not being calculable (i.e., not “converging”). This is true for the present data. Practically therefore, following the recommendation of Barr et al. (under review), the statistical models presented here include a number of control effects, and random effects by subject for only the experimental effect of response-distractor congruency (the Distractor conditions), and for one control effect that addresses the fact that subjects vary as to whether their RTs tended to slow down or speed up over the course of the experiment (this factor will be discussed in more detail in section 3.2.2.1 below).

### **3.2.2.2. Results of statistical models**

The response-time data from this experiment were analyzed with two statistical models. The first mixed-effect model included only control effects. The second model included the same control effects plus the experimental effect. The significance of the experimental effects will be shown in three ways. First and most importantly, the individual effects of congruency will be shown to be significant within the model that includes the experimental effect. Second, the two models themselves will be compared statistically, showing that the model that includes the experimental effect accounts for significantly more variance in the data than the model that does not include the experimental effects. Third, *p* values will be calculated for the congruency effects using Monte Carlo Markov chain sampling, though this test should be interpreted cautiously (as will be discussed in section 3.2.2.2). Response-time data were analyzed using R (R development core team, 2010), using the `lme4` package (Bates, 2005; Bates & Maechler, 2009) for the mixed-effects modeling. Data were graphically presented using the `sciplot` package for R (Morales, 2011).

#### **3.2.2.2.1. Control effects**

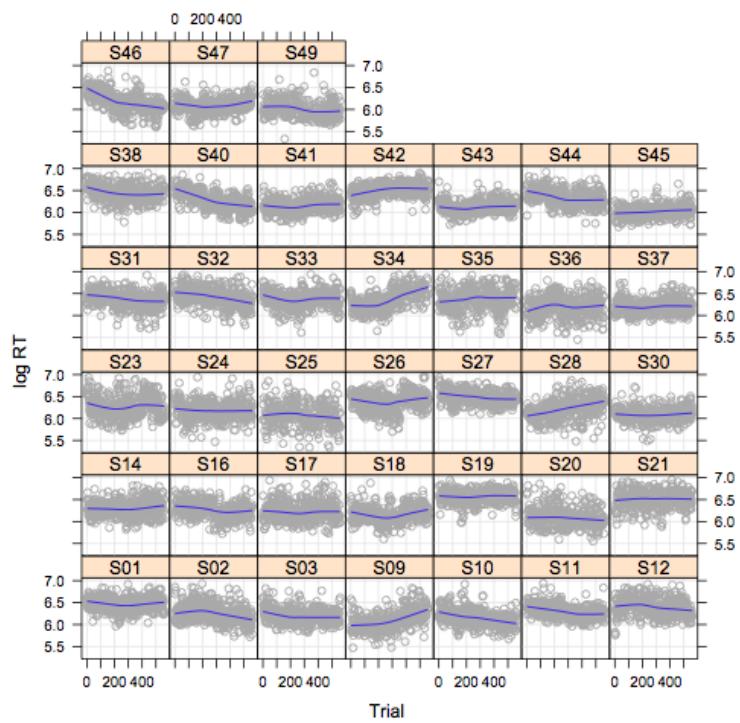
This section presents the results from the statistical model that includes several control factors that have been shown to have an influence on RTs. Baayen, Davidson, & Bates (2008) report that in lexical decision tasks, the RT of the preceding trial is a very reliable predictor of RT on the current trial: if subject took a relatively long time

to respond on trial<sub>x</sub>, he or she is likely to reply relatively slowly on trial<sub>x+1</sub>. Whether the answer to the previous trial was correct also influences the RT of the current trial: when the subject makes a mistake, he or she is likely to respond more slowly on the next trial (see Dutilh et al., 2012, for discussion). Subjects also respond more quickly on a trial if the response is the same as the response for the previous trial, which is most straightforwardly interpreted as a selection effect (Galantucci et al., 2009; Kerzel & Bekkering, 2000). Kerzel and Bekkering (2000) and Galantucci et al. (2009) found a strong main effect of SOA on RTs in this experimental task, with longer RTs at higher SOAs.

Item—i.e., the syllable that the subject uttered, which in this experiment was *da*, *ta*, *ga*, or *ka*—was included in the model as a random effect so that any differences in RTs attributable to the response were accounted for independently from the other effects to be tested. Models with by-item slopes for SOA, etc., did not converge, so only a by-item intercept was included.

Mixed modeling also allows longitudinal effects that unfold over the course of the experiment to be included in accounting for variance in the response-time data. Such longitudinal effects may include habituation and fatigue. Subjects were a random factor in the experimental design and did not demonstrate uniform behavior. Figure 8 shows the RTs by subject over the course of the two experimental blocks. Figure 8 reveals several facts about the data. First, the average RT changed from subject to subject. For example, the RTs of S19 were visibly slower than those of S20. Second,

some subjects exhibited obvious longitudinal changes in RT (e.g., S46 and S34) while others did not (e.g., S17, S24). Lastly, of those that did show longitudinal effects, some responded faster as the experiment progressed (e.g., S10, S46), possibly due to habituation, while others responded more slowly as the experiment progressed (e.g., S09, S34), possibly due to fatigue.



**Figure 8. Log response times over the course of the experiment (both blocks) by subject.**

All of the control effects (random and fixed) included in the model are summarized in Table IV, along with the expected result for each. Trial and Previous Trial RT were defined as interval scales, as was SOA, though SOA only had three

possible values. Previous Trial Correct and Same Response as Previous Trial were logical predictors. Item was a four-level categorical effect.

Before proceeding to the results from the model, one technical issue should be addressed with the by-Subject Trial random effect. Each effect has an intercept and a slope associated with it, and these are assumed to be independent. Since trial numbers are bounded on the left and cannot be less than zero, changes in slope across Subject

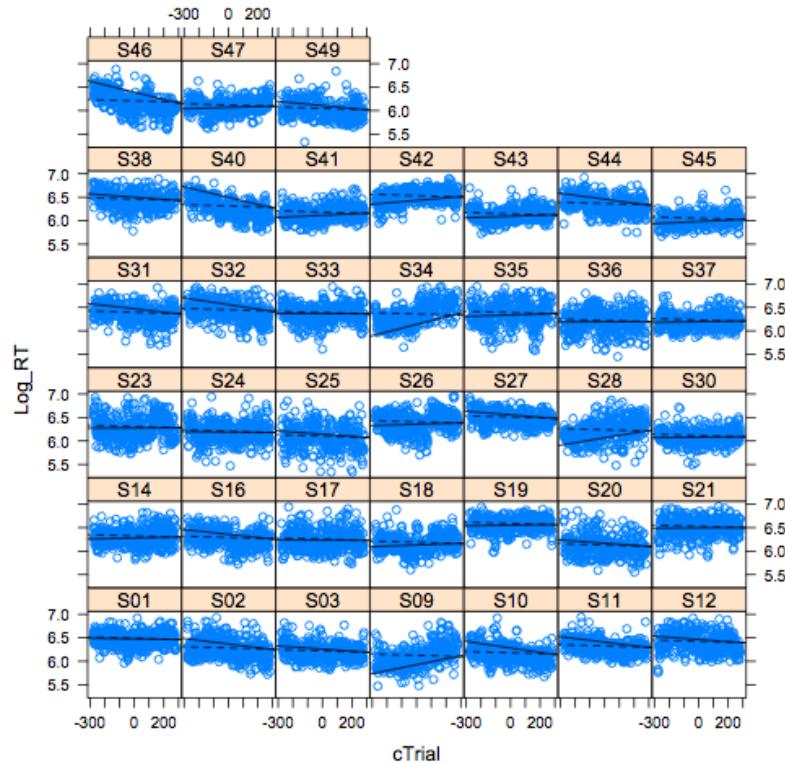
**Table IV. Control random and fixed effects on RTs included in the model.**

| Effect                           | Levels ( <i>values</i> )                        | Type   | Expectation  |
|----------------------------------|---|--------|--|
| Item                             | 4 ( <i>da, ta, ga, ka</i> )                     | Random | None   |
| Trial                            | interval  | Fixed  | None   |
| Trial * Subject                  |   | Random | Effects of cTrial will vary by Subject.  |
| SOA                              | 3 values (100, 200, 300 ms) treated as interval | Fixed  | Positive correlation with RT.<br>Longer SOAs yield longer RTs.   |
| Previous Trial RT                | interval  | Fixed  | Positive correlation with RT.  |
| Previous Trial Correct?          | 2 values ( <i>I = Yes, 0 = No</i> )             | Fixed  | Negative correlation with RT.<br>Subjects should reply slower if previous trial was incorrect.   |
| Same response as Previous Trial? | 2 values ( <i>I = Yes, 0 = No</i> )             | Fixed  | Negative correlation with RT.<br>Subjects should reply slower if the response on a given trial was the same as the response on the previous trial. |

may correlate with changes in intercept (see Baayen, 2008, p. 254, Figure 7.5). Baayen (2008) recommends “centering” the Trial values to avoid this problem. This was achieved by subtracting the mean of the Trial values from each Trial value. This allowed the slopes of the by-Subject Trial effects to vary without affecting the intercepts. The fixed effect of Trial and the by-Subject random effect of Trial was measured based on centered Trial values. Figure 9 compares the intercepts and slopes assigned when Trial was modeled as a fixed effect only, with a model in which in addition to Trial being a fixed effect the interaction of Trial and Subject was also modeled. When only Trial was included in the model as a fixed effect, the intercept could vary with subject, but the slope could not. Since these subjects behaved differently over the course of the experiment, the Trial and Subject interaction was included in the model.

Centering predictor values can also have the benefit of reducing the collinearity between predictors (though see Echambadi & Hess, 2007, for further discussion). Predictors in a mixed-effect model should ideally not be correlated at all, and highly correlated predictors can cause serious problems for mixed-effect modeling (Belsley, Kuh, & Welsch, 2004). One way to reduce spurious collinearity between effects is to center all interval predictors (see Gelman & Hill, 2007, Chapter 4, for discussion). Centering predictor values (or performing any linear transformation on them) has no effect on the estimates, slopes, or significance of the predictors (Gelman & Hill, 2007). Therefore, all of the interval effects in the model shown in Table IV

were centered, using functions provided in Austin Frank's `mer-utils` and `regression-utils` (available <http://github.com/aufrank/R-hacks> as of April 2012), which were also used to calculate the collinearity measures presented below.



**Figure 9. Comparison of the intercept and slope assigned when centered Trial is modeled only as a fixed effect (dotted line) with the interaction of Trial and Subject also modeled as a random effect (solid line).**

Table V shows the results of the model that contained only the control effects described above. The values in the Estimate column indicate the coefficients for the Intercept and the slope associated with each fixed effect, that is, the amount by which the effect is expected to change the value of the dependent variable from the Intercept.

The value shown in the “ms” column is the estimate converted from log-space back to millisecond space according to the formula in (1), where X is a given effect.

$$(1) \quad X_{ms} = e^{(Estimate_{Intercept} + Estimate_{Xlog})} - e^{(Estimate_{Intercept})}$$

One challenge in interpreting linear mixed-effects models is that it is currently unclear how to calculate the correct degrees of freedom (Bates, 2005). Baayen and colleagues (Baayen, 2008; Baayen et al., 2008) advocate a threshold  $t$  value of 2, where an absolute  $t$  value greater than 2 indicates a significant effect. This was the criterion adopted here to indicate significance. Since this model investigated control effects only, no further measures were taken to determine the significance of the effects. In the model in the next section that tested for the significance of the experimental effects, additional analyses were performed, as described at the beginning of this section.

The `lme4` package in R does not list the proportion of the variance ( $R^2$ ) accounted for by the model, because multiple sources of variance are modeled at once, including both random and fixed effects. The amount of variance in the data that the entire model accounts for can be calculated as the correlation between the observed variance in the data and the fitted values of the model. The  $R^2$  values reported here were calculated this way. Collinearity measures include: the condition number ( $K$ ), which when below 6 indicates virtually no collinearity (Baayen, 2008, p. 182); the Variance Inflation Factor (VIF), which should be below 5 (Belsley et al., 2004); and

the maximal correlation between any two fixed effects, for which only absolute values greater than 0.4 are generally viewed as acceptable requiring further explanation.

**Table V. Results of a linear mixed-effects model of control effects on Log RT in the voice experiment. “c.” indicates that the effect values were centered. The slope coefficients for the fixed effects are shown in the “Estimate” column, expressed in log RT, and indicate the magnitude and direction of the effect compared to the baseline condition, shown in the Intercept row. A millisecond equivalent of the Estimate for each fixed effect is shown in the “ms” column. The effect slope for SOA was extremely small considering the obvious effect of SOA that can be seen in Figure 5C and D. This is due to the fact that SOA was coded as a continuous variable, but it only had three possible values (100, 200 or 300 ms). The slope indicates the adjustment to the Intercept for each incremental unit of the effect, which in the case of SOA was 1 ms, so the actual adjustment for SOA is 100 times the slope shown and is indicated in parentheses to the right of the ms value. Significance of each effect is indicated in the “t value” column.  $|t| > 2$  indicates a significant effect.**

```

Linear mixed model fit by REML
Formula: Log_RT ~ c.(SOA) + c.(Prev_correct) + c.(PrevSame) +
          c.(PrevRT_Log) + c.Trial + (1 + c.Trial | Subject) +
          (1 | Item)

      AIC      BIC    logLik   deviance   REMLdev
 -10137  -10052     5079     -10229     -10159

Random effects:
 Groups   Name        Variance   Std.Dev.
 Subject  (Intercept) 0.01859   0.13633341
           c.Trial     0.00000   0.00030865
 Item     (Intercept) 0.00033   0.01817121
 Residual            0.03064   0.17503573

Number of observations: 16337, groups: Subject, 38; Item, 4

Fixed effects:
              Estimate     ms   Std. Error   t value
(Intercept)    6.26800   527   0.02395   261.72
c.(SOA)        0.00044   0(23)  0.00002   25.91
c.(Prev_correct) -0.08012  -41   0.01156   -6.93
c.(PrevSame)    -0.02259  -12   0.00275   -8.21
c.(PrevRT_Log)   0.04193   23   0.00315   13.33
c.Trial        -0.00009   0    0.00005   -1.72

R² = 0.4501,  K = 1.149,  VIF = 1.005,  max corr = -0.055

```

The fixed effects were consistent with expectations (where they existed) and all were significant, with the exception of Trial. The SOA slope coefficient was positive, meaning RTs got significantly longer as SOA increased. The slope coefficient for Correctness of Previous Trial (Prev\_correct) was negative, meaning an incorrect reply on the previous trial resulted in larger RTs (since Prev\_correct was coded as 0 if the response on the previous trial was incorrect and 1 if it was correct). The slope coefficient for Previous Trial RT (PrevLogRT) was positive, meaning that the RT on a given trial was a strong predictor of the RT on the next trial, i.e., longer RTs begot longer subsequent RTs. The slope coefficient for whether the response was the same as the response of the previous trial (PrevSame) was negative, meaning people responded more quickly on a trial if the response was the same as the response of the previous trial (PrevSame was coded so that 0 means the previous response was different and 1 means it was the same). The model accounted for 45.01% of the variance in the response-time data. All of the collinearity measures were exceedingly low, indicating that the predictors of the model were not correlated.

### **3.2.2.2.2. Experimental effects**

This section presents the results of an expanded mixed model of the response-time data from Experiment 1. This model included all of the effects in the model presented above in section 3.2.2.2.1 plus the Distractor, which had three levels: Tone, Congruent, or Incongruent (shown in Table VI). The Tone Distractor refers to those

trials where the subjects heard a tone. Congruent means that the response shared voicing with the distractor (but differed in articulator), whereas Incongruent means that the response differed in voicing from the distractor (as well as differing in articulator). Each of the four Items occurred with Tone, Congruent, and Incongruent Distractors (see Table I). A random effect was included to account for subject-specific differences in the various Distractor conditions (per Barr et al., under review).

**Table VI. Experimental effects on RT added into the model.**

| Effect     | Levels ( <i>values</i> ) | Type   | Expectation   |
|------------|--------------------------|--------|---|
| Distractor | 3                        | Fixed  | <i>Hypothesis:</i><br><i>(Tone, Congruent,</i><br><i>Incongruent)</i> RTs in Congruent trials should be shorter than in Incongruent trials. |
| Distractor |                          | Random | None: accommodates within-subject differences for the Distractor conditions   |
| * Subject  |                          |        |   |

A new statistical model was created that included all of the control effects from the previous model (Table VI), plus Distractor. The results are shown in Table VII. In this model as in the previous model, the control fixed effects of SOA, Correctness of Previous Trial (Prev\_Correct), Previous Trial RT (PrevLogRT), and whether the Previous response was the same (PrevSame) were significant, and qualitatively the same as the previous model. Trial was not significant.

The Distractor effect, unlike all of the others, was not scalar but categorical. The `lme4` package treats each value of such factors as its own fixed effect. The `lme4` package

provides slope coefficients for  $N - 1$  of the  $N$  values for nominal-scale fixed effects, where  $N$  in this case was 3. These are to be interpreted as compared to the baseline value, which is the case that is not shown. In this particular model, the Congruent Distractor is not listed, which means that the slope coefficients of the two Distractors listed are relative to the coefficient of the Congruent Distractor slope, included in the Intercept. The negative slope coefficient of the Tone Distractor (slope coefficient = -0.01488) shows that RTs were shorter when there was a Tone Distractor than when there was a Congruent Distractor, and the  $t$ -value ( $t = -4.09$ ) indicates that this effect was significant. The positive slope coefficient of the Incongruent Distractor (slope coefficient = 0.00755) shows that RTs were longer when there was an Incongruent Distractor than when there was a Congruent Distractor, and significantly so ( $t = 2.22$ ). This model that included the experimental effects accounted for 45.21% of the variance in the data.<sup>4</sup>

The collinearity measures of  $K$  and VIF were very low in this model as well. The maximum correlation between any two fixed effects was 0.465. This was the correlation between two Distractor conditions: Tone and Incongruent. Given these were two mutually-exclusive sets of trials because they represent two levels of the same condition, it is difficult to attribute any particular meaning to the correlation

---

<sup>4</sup> Ideally, the model would also have included a term for the interaction between Distractor and SOA. Unfortunately, `lme4` would not compute a model that included that interaction, for reasons that remain unclear. Impressionistically, the results shown in Figure 5B show a consistent pattern of Distractor by SOA, with no numerical reversals in any of the SOAs, so it is assumed that that interaction would not have been significant (though this remains to be tested statistically).

between the two effects. The next-highest correlation between two fixed effects was 0.109, so it seems safe to conclude that collinearity was not a problem in this model.

**Table VII. Results of a linear mixed-effects model including the experimental effects on Log RT in the voice experiment (Experiment 1). “c.” indicates that the effect values were centered. The slope coefficients for the fixed effects are shown in the “Estimate” column, expressed in log RT, and indicate the magnitude and direction of the effect compared to the baseline condition, shown in the Intercept row. A millisecond equivalent of the Estimate for each fixed effect is shown in the “ms” column. Significance of the effect is indicated in the “t value” column, where  $|t| > 2$  indicates a significant effect. The effect of the Incongruent Distractor condition compared to the Congruent Distractor condition (included in the Intercept) is highlighted in the box.**

```

Linear mixed model fit by REML
Formula: Log_RT ~ c.(SOA) + c.(Prev_correct) + c.(PrevSame) +
         c.(PrevRT_Log) + c.Trial + Distractor + (Distractor +
         c.Trial | Subject) + (1 | Item)

      AIC      BIC    logLik   deviance   REMLdev
 -10155  -10001     5098     -10285     -10195

Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 0.01917  0.1384370
          Distractor:Incongruent 0.00001  0.0030545
          Distractor:Tone       0.00008  0.0088748
          c.Trial                0.00000  0.0003084
 Item     (Intercept) 0.00033  0.0181180
 Residual                         0.03053  0.1747359

Number of observations: 16337, groups: Subject, 38; Item, 4

Fixed effects:
            Estimate    ms   Std. Error  t value
(Intercept) 6.27000  528    0.02433  257.71
c.(SOA)      0.00044  0(23)  0.00002  25.95
c.(Prev_correct) -0.08069 -41    0.01154  -6.99
c.(PrevSame)  -0.02261 -12    0.00275  -8.23
c.(PrevRT_Log) 0.04196  23    0.00314  13.36
c.Trial      -0.00009  0     0.00005  -1.72
Distractor: Incongruent 0.00755  4     0.00340  2.22
Distractor: Tone      -0.01488 -8     0.00364  -4.09

R2 = 0.4521,  K = 3.179,  VIF = 1.308,  max corr = 0.465

```

As discussed earlier, there is still a lack of clarity in the statistics literature as to the best method for establishing the significance of an effect when using linear mixed-effects models. Three methods are shown here. The first is shown above, with an absolute t-value greater than 2 indicating significance. The Congruent condition is significant using that criterion.

**Table VIII. ANOVA comparison of the model with control effects only with the model that included the experimental effects of Distractor congruency. The model with the experimental effects accounted for significantly more of the variability in the data than the model without the linguistic effects, as indicated in the “Pr(>Chisq)” column.**

|                | Df | AIC      | BIC      | logLik | Chisq  | ChiDf | Pr(>Chisq)  |
|----------------|----|----------|----------|--------|--------|-------|-------------|
| voice.Controls | 11 | -10207.2 | -10122.5 | 5114.6 |        |       |             |
| voice.Distrs   | 20 | -10244.8 | -10090.8 | 5142.4 | 55.647 | 9     | < 0.0000*** |

The second method is to use ANOVA to compare two models that differ only in the presence or absence of the experimental effect. If the model with the experimental effects accounts for more variance in the data and the result of the ANOVA shows a significant difference between the models, this is further evidence that the experimental effects were significant. The model that had only control effects captured 99.6% (0.4501 / 0.4521) of the variance that the full model captured, meaning 0.4% of the total variance that was modeled was attributable to the experimental effects of interest. Table VIII shows the results of an ANOVA

comparing the two models, demonstrating that the additional variability accounted for by the model with the experimental factors, though small, was significant.

**Table IX. Results of a linear mixed-effects model including linguistic effects on Log RT in the voice experiment, without subject-specific slopes for Trial or Distractor.**

```

Linear mixed model fit by REML
Formula: Log_RT ~ c.(SOA) + c.(Prev_correct) + c.(PrevSame) +
          c.(PrevRT_Log) + c.Trial + Distractor +
          (1 | Subject) + (1 | Item)

AIC    BIC  logLik  deviance  REMLdev
-9116 -9031   4569      -9229     -9138

Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 0.01769960 0.133040
 Item     (Intercept) 0.00033572 0.018323
 Residual            0.03285398 0.181257
Number of observations: 16337, groups: Subject, 38; Item, 4

Fixed effects:
              Estimate    ms  Std. Error t value
(Intercept)  6.62700  528    0.02357  265.98
c.(SOA)       0.00046  0(24)  0.00002  25.87
c.(Prev_correct) -0.08820 -45    0.01194  -7.39
c.(PrevSame)  -0.02223 -12    0.00285  -7.81
c.(PrevRT_Log) 0.05834  32    0.00322  18.14
c.Trial      -0.00011  0     0.00002  -6.39
Distractor: incongruent 0.00792  4    0.00349  2.27
Distractor: tone      -0.01506 -8    0.00346  -4.35

R² = 0.4090,  K = 3.179,  VIF = 1.336,  max corr = 0.501

```

A third way to establish significance of effects is to calculate  $p$  values using Markov chain Monte Carlo (MCMC) sampling (though see discussion at the end of this sub-section). A slight adjustment to the present models was necessary before doing so, as the current implementation of `lme4` cannot perform MCMC sampling on models that include interactions between random and fixed effects. The present models included two such terms (Distractor + Trial | Subject). In order to use the

MCMC sampling, a new model was created that included the experimental effects but removed these terms, with Subject and Item as a random effects. The results of the model are shown in Table IX.

While very similar overall, there were a few differences to note in comparing this model (“NoI|S” for short) with the model that included the subject-specific Trial slopes (“Full” for short). The NoI|S model accounted for less variability than the Full model (40.90% vs. 45.21%), and the fixed effect of Trial was significant (assuming the  $|t| > 2$  criterion). However, the  $t$  value for the Incongruent Distractor case only changed by 0.05 between the models (with no change in the back-transformed slope expressed in milliseconds). This motivates considering the MCMC sampling results for the NoI|S model, which are shown in Table X.

**Table X. Results of Markov chain Monte Carlo sampling (with 10000 samples) for the voice experiment. Significance determined based on the  $p$  value calculated by MCMC sampling are shown in the “ $p_{MCMC}$ ” column. The box around the Incongruent vs. Congruent Distractor conditions show that the effect of congruency was significant with a Bonferroni-corrected alpha of  $0.05/2 = 0.025$ .**

|                    | Estimate | MCMCmean | HPD95lower | HPD95upper | $p_{MCMC}$ |
|--------------------|----------|----------|------------|------------|------------|
| (Intercept)        | 6.2700   | 6.2699   | 6.2001     | 6.3385     | 0.0001     |
| c.(SOA)            | 0.0005   | 0.0005   | 0.0004     | 0.0005     | 0.0001     |
| c.(Prev_correct)   | -0.0882  | -0.0881  | -0.1116    | -0.0651    | 0.0001     |
| c.(PrevSame)       | -0.0222  | -0.0222  | -0.0278    | -0.0167    | 0.0001     |
| c.(PrevRT_Log)     | 0.0583   | 0.0586   | 0.0523     | 0.0650     | 0.0001     |
| c.Trial            | -0.0001  | -0.0001  | -0.0001    | -0.0001    | 0.0001     |
| <b>Distractor:</b> |          |          |            |            |            |
| incongruent        | 0.0079   | 0.0079   | 0.0014     | 0.0149     | 0.0226     |
| <b>Distractor:</b> |          |          |            |            |            |
| tone               | -0.0151  | -0.0150  | -0.0219    | -0.0086    | 0.0002     |

Since there were multiple comparisons within the Distractor condition, the alpha of 0.05 should be corrected by dividing it by the number of comparisons. The Bonferroni-corrected alpha was  $0.05/2 = 0.025$ . The results of the MCMC sampling (with 10000 samples) show that the results were qualitatively the same whether using this  $p$  value or using the method in the Full model above based on the absolute value of the  $t$ -value being greater than 2. Three different tests of significance all show that the effect of congruency remained: RTs on trials with Incongruent Distractors were significantly longer than those with Congruent Distractors.

It should be emphasized, however, that this significance test the using the MCMC sampling should be interpreted cautiously (or potentially ignored should the differences between the models reported in Table VII and Table IX be material), since it does not test the actual model reported in Table VII. Nevertheless, the significance of the Congruency effects is sufficiently established by the  $t$  values of the model reported in Table VII, and the significant ANOVA model comparison reported in Table VIII.

### **3.2.3. Discussion**

The control fixed effects each had a significant influence on the RTs in the task, in the directions expected. Subjects were slower to respond if they made a mistake on the previous trial, and quicker if the response on a given trial was the same as the previous trial. Larger SOAs resulted in longer RTs, which is consistent with

what has been found for this task by Kerzel & Bekkering (2000) and Galantucci, Fowler, & Goldstein (2009). The RT of the previous trial was a good predictor of the RT for the current trial.

The linguistic factor of Distractor congruency was also significant, in support of the hypothesis that voicing plays a role independent of articulator in the link between perception and production. This is consistent with the findings of Gordon and Meyer (1984), though the present results are more clearly identifiable as perceptuo-motor effects. When subjects heard speech Distractors in the present experiment, it always differed in articulator from what their response was supposed to be. Congruent Distractors differed in articulator but matched in voicing, whereas Incongruent Distractors different in both articulator and voicing. These data suggest that whenever there was a speech distractor, the mismatch in articulator resulted in a slowdown in RTs, since RTs were shorter when they heard a tone distractor than when they heard a speech distractor. However, the penalty on RTs was less in the Congruent case than in the Incongruent case. This result supports the hypothesis that RTs are sensitive to the temporal property of voicing (independently from the spatial property of articulator), and provides the first clear experimental evidence for a role of voicing independent of articulator in the perception-production link.

### **3.3. Articulator RT experiment (Experiment 2)**

This experiment tested whether a perceptuo-motor effect of articulator could be identified. If such an effect is present, then RTs on trials when the response and distractor are speech sounds made with the same articulator should be shorter than trials where they are sounds made with different articulators.

#### **3.3.1. Methods**

The methods for this experiment were the same as those for the voicing experiment described above in section 3.2.1, except as noted below. All materials and procedures were approved by the Institutional Review Board of New York University (the University Committee on Activities Involving Human Subjects).

##### **3.3.1.1. Subjects**

40 subjects were recruited from the subject pool of the New York University Department of Psychology, and received credit in fulfillment of class requirements. All subjects identified themselves as native speakers of American English, and having no speech or hearing impairments. Data from five subjects had to be excluded for various reasons: two subjects misunderstood the task and consistently made incorrect responses on all trials, e-Prime generated large timing errors for a large number of trials for two subjects, and one subject was chewing gum during the experiment. Data from the remaining 35 subjects (28 female, 7 male) were analyzed. All subjects gave informed consent before participating in the experiment.

### 3.3.1.2. Procedure

The procedure for this experiment was the same as for the experiment 1 as described in section 3.2.1.2, except that there were four blocks in this experiment instead of two. The response-distractor pairs for each block are shown in Table XI. Any general facilitative effect of articulator should be present across multiple articulators. Therefore, responses with three different articulators were included in the design (lips, tongue tip, and tongue back). An effect of articulator should also be found for consonants other than oral stops. A block with nasal responses was therefore also included.

**Table XI. Response-distractor pairs for the articulator experiment. A green check mark ( $\checkmark$ ) indicates that the response and distractor shared articulator (Congruent condition), a red X indicates that they differed in articulator (Incongruent condition). Dots show the tone and no distractor conditions, which were also included for each block. An empty cell indicates that response-distractor pair was not part of that block.**

|         |    | Distractor |              |              |              |      |      |
|---------|----|------------|--------------|--------------|--------------|------|------|
|         |    | Response   | ba           | da           | ga           | Tone | none |
| Block 1 | pa |            | $\checkmark$ | X            |              | •    | •    |
|         | ta |            | X            | $\checkmark$ |              | •    | •    |
| Block 2 | ma |            | $\checkmark$ | X            |              | •    | •    |
|         | na |            | X            | $\checkmark$ |              | •    | •    |
| Block 3 | pa |            | $\checkmark$ |              | X            | •    | •    |
|         | ka |            | X            |              | $\checkmark$ | •    | •    |
| Block 4 | ta |            |              | $\checkmark$ | X            | •    | •    |
|         | ka |            |              | X            | $\checkmark$ | •    | •    |

Responses were voiceless oral stops in three blocks (1, 3, and 4) or nasal stops in one block (block 2), while distractor syllables were always voiced oral stops. Therefore in three blocks (1, 3, and 4), responses and distractors always differed in voicing. In block 2, responses and distractors were both voiced, but responses were nasal whereas distractors were oral. In all blocks, Incongruent cases differed in articulator (e.g., *ba-ta*), whereas Congruent cases shared articulator (e.g., *ba-pa*).

The audio distractors were presented with an SOA of 100, 200, or 300 ms after the onset of the visual cue. Each response-distractor combination was presented 14 times at each of the three SOAs, yielding 252 trials per block. There were 28 trials in each block (14 per response) where the subjects did not hear a distractor, bringing the total to 280 trials per block, 1120 trials per experiment. Trials were pseudo-randomized within block. The order of presentation of the blocks was pseudo-randomized by subject, as was whether the subject saw == or # # for a given response in a given block. The entire experiment including the two practice blocks and four experimental blocks lasted about 50 minutes per subject. The equipment setup was identical to the setup described in section 3.2.1.3.

### 3.3.1.3. Stimuli

There were three distractor syllables used in this experiment (*ba*, *da*, *ga*), plus the tone distractor. The *ba* and tone sound files were the same ones used in the voicing experiment described in section 3.2.1.4. The *da* and *ga* distractor stimulus were

recorded by the same speaker in the same recording session as the stimuli in the voicing experiment.

**Table XII. Properties of the distractor stimuli in the articulator experiment.**

| stimulus  | VOT   | vowel duration | format transitions   | F0 (Hz) range | F1 (Hz) end | F2 (Hz) end |
|-----------|-------|----------------|----------------------|---------------|-------------|-------------|
| <i>ba</i> | 0 ms  | 150 ms         | ~27 ms<br>(5 pulses) | 179-193       | 858         | 1431        |
| <i>da</i> | 9 ms  | 141 ms         | ~49 ms<br>(8 pulses) | 179-195       | 834         | 1460        |
| <i>ga</i> | 15 ms | 137 ms         | ~50 ms<br>(9 pulses) | 175-202       | 909         | 1487        |

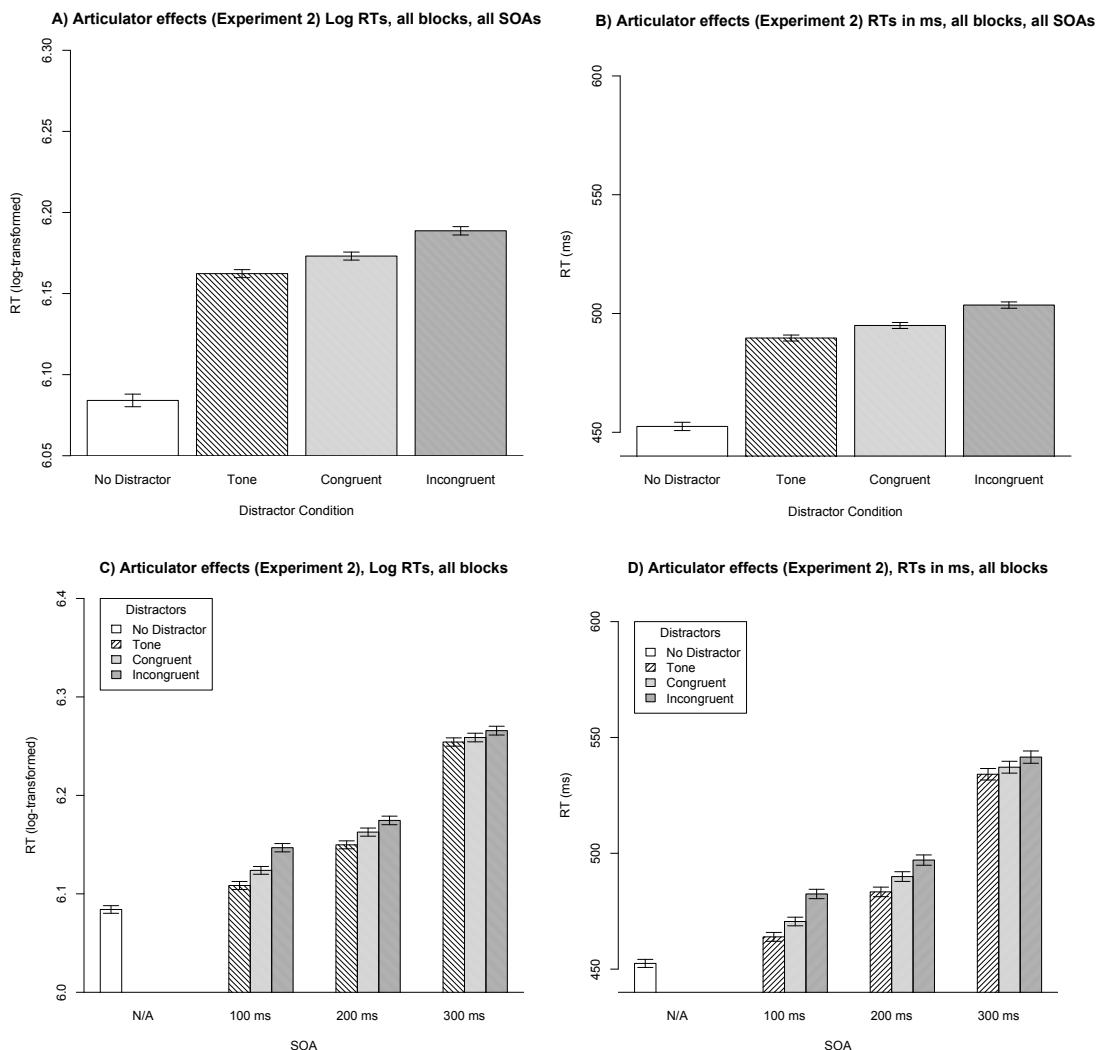
Table XII shows the properties of the syllable distractor stimuli used in the articulator experiment. The values for the *ba* stimulus are repeated from Table II for convenience of comparison.

### 3.3.2. Results

35 subjects yielded 39200 trials. Two types of erroneous data were excluded based on the same criteria as in the voicing experiment: trials where e-Prime introduced timing errors resulting in an actual SOA differing by more than 30 ms from the specified SOA, and trials where the subject did not produce the correct syllable based on the judgment of the experimenter. Unlike the voicing experiment, any incorrect response on a trial was excluded from analyses of this experiment. That is, if

the subject was supposed to reply *ta* but replied *pa* in the *ta-pa* block, that trial was excluded. As in the voicing experiment, certain responses were excluded from analysis based on RT. Trials were excluded if the subject's response started earlier than 100 ms into the playing of the audio distractor. It does not seem reasonable to expect the distractor to evoke perceptuo-motor effects if the subject starts a production before he or she has had time to perceive the audio distractor. Trials were excluded from analysis if the RT was greater than 750 ms after the presentation of the distractor, on the assumption that the subject was inattentive on that trial. Trials with no distractor were excluded from analysis due to the complete conflation of the No Distractor and "N/A" SOA conditions. 13735 trials (35.0% of the total) were excluded based on these criteria, leaving 25465 trials to be included in the analyses. The average RT of the remaining trials was 496 ms.

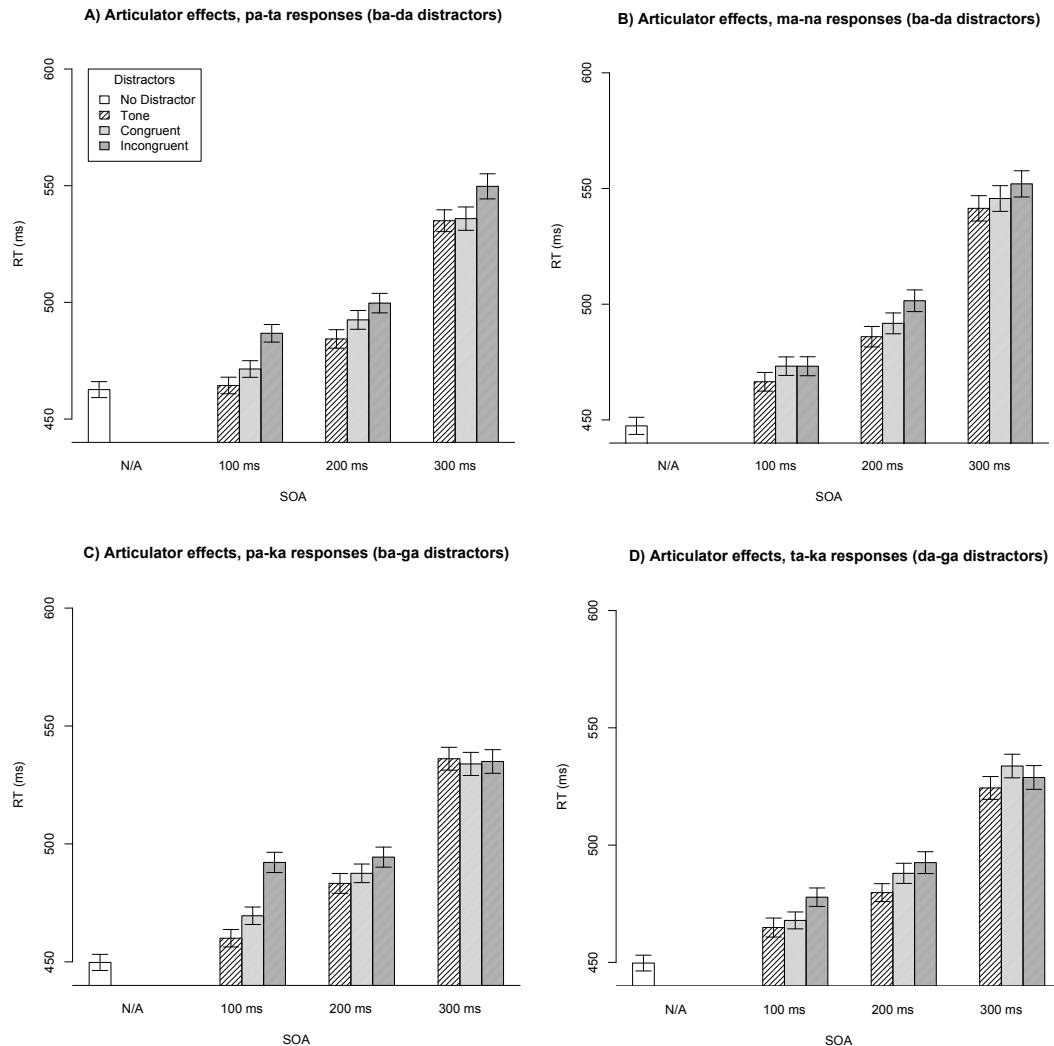
The methods used for data labeling and statistical analysis were identical in this experiment to the methods of the voicing experiment as described in sections 3.2.2. RTs were log-transformed.



**Figure 10. Mean RTs for the articulator experiment (Experiment 2).** A) shows the means of the log-transformed RTs by Distractor across all blocks and SOAs, which were subject to statistical analyses while B) shows the mean RTs in milliseconds. Error bars represent 95% confidence intervals. The No Distractor condition was not included in the statistical analyses, but is shown here for comparison. C) shows the log-transformed RTs in A) broken down by SOA. D) shows the ms RTs in B) broken down by SOA.

Mean RTs are shown in Figure 10, with Figure 10A showing the mean of the log-transformed RTs, which were submitted to statistical analyses, while Figure 10B shows the mean RTs in milliseconds for a more intuitive comparison of the response-

time patterns. Figure 10 shows that the hypothesis that RTs in the Congruent condition should be shorter than in the Incongruent condition was supported. The statistical analyses below show that this difference was significant.



**Figure 11. Mean RTs in ms for Experiment 2, by block. A) pa-ta responses, B) ma-na responses, C) pa-ka responses, D) and ta-ka responses. Error bars represent 95% confidence intervals.**

Figure 11 shows the RT patterns (in milliseconds) within each experimental block. The overall pattern shown in Figure 10 is generally replicated in each block. Only in block 3 at SOA is the mean RT in the Incongruent condition numerically less than in the Congruent condition.

### 3.3.2.1. Control effects

The same control effects were modeled as in the voicing experiment (see Table IV), with two minor changes. First, in this experiment there were five Items: *pa*, *ta*, *ka*, *ma*, and *na*. Second, the model would not converge when all of the predictor values were centered. A model that converged was obtainable by simply not centering just one of the logical predictors (whether the previous response was correct).

Table XIII shows the results of the model that contains only the control effects. The  $R^2$  and collinearity values were calculated as discussed in section 3.2.2.

**Table XIII. Results of a linear mixed-effects model of control effects on Log RT in the articulator experiment.**

```

Linear mixed model fit by REML
Formula: Log_RT ~ c.(SOA) + Prev_correct + c.(PrevSame) +
         c.(PrevRT_Log) + c.Trial + (1 + c.Trial | Subject) +
         (1 | Item)

      AIC      BIC  logLik deviance  REMLdev
 -11270  -11180    5646     -11365    -11292

Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 0.01813  0.13464
           c.Trial     0.00000  0.00008
 Item     (Intercept) 0.00000  0.00136
 Residual            0.03704  0.19246

Number of observations: 25465, groups: Subject, 35; Item, 5

Fixed effects:
             Estimate    ms  Std. Error  t value
(Intercept) 6.22700  506    0.02327  267.58
c.(SOA)      0.00048  0(24)  0.00002  30.94
Prev_correct -0.09553 -46    0.00812  -11.77
c.(PrevSame) -0.02947 -15    0.00312  -9.45
c.(PrevRT_Log) 0.04134  21    0.00337  12.25
c.Trial      -0.00002  0     0.00001  -1.44

R² = 0.3083,  K = 10.406,  VIF = 1.085,  max corr = -0.274

```

The fixed effects were all significant, with the exception of cTrial. All results were consistent with expectations, and very similar to the results of the similar model reported for Experiment 1 (Table V). The model accounted for 30.83% of the variance in the response-time data. The collinearity values were well within acceptable limits.

### 3.3.2.2. Experimental effects

A second model was created that added the experimental effect of Distractor, which had three levels: Tone, Congruent, or Incongruent. “Congruent” meant that the

response shared articulator with the Distractor, whereas Incongruent meant that the response differed in articulator from the Distractor. Each of the five Items occurred with Tone, Congruent, and Incongruent Distractors. A by-subject random term was included in the model, as shown in Table VI for experiment 1. The hypothesis tested was that RTs on trials where the response and Distractor shared articulator

**Table XIV. Results of a linear mixed-effects model including experimental effects on Log RT in the articulator experiment. The Incongruent Distractor condition is highlighted in the box, showing its RTs were significantly longer than the Congruent Distractor condition, included in the Intercept as the baseline condition.**

| Linear mixed model fit by REML  |                        |          |            |         |  |
|---|------------------------|----------|------------|---------|--|
| Formula: Log_RT ~ c.(SOA) + Prev_correct + c.(PrevSame) + c.(PrevRT_Log) + c.Trial + Distractor + (Distractor + c.Trial   Subject) + (1   Item) |                        |          |            |         |  |
| AIC   | BIC                    | logLik   | deviance   | REMLdev |  |
| -11318  | -11155                 | 5679     | -11451     | -11358  |  |
| <b>Random effects:</b>  |                        |          |            |         |  |
| Groups  | Name                   | Variance | Std.Dev.   |         |  |
| Subject   | (Intercept)            | 0.01810  | 0.13451975 |         |  |
|   | Distractor:Incongruent | 0.00004  | 0.00613529 |         |  |
|   | Distractor:Tone        | 0.00003  | 0.00534228 |         |  |
|   | c.Trial                | 0.00000  | 0.00008115 |         |  |
| Item  | (Intercept)            | 0.00000  | 0.00143086 |         |  |
| Residual  |                        | 0.03690  | 0.19209908 |         |  |
| Number of observations: 25465, groups: Subject, 35; Item, 5   |                        |          |            |         |  |
| <b>Fixed effects:</b>   |                        |          |            |         |  |
|   | Estimate               | ms       | Std. Error | t value |  |
| (Intercept)   | 6.22600                | 506      | 0.02331    | 267.10  |  |
| c.(SOA)   | 0.00048                | 0(24)    | 0.00002    | 30.98   |  |
| Prev_correct  | -0.09579               | -46      | 0.00810    | -11.83  |  |
| c.(PrevSame)  | -0.02941               | -15      | 0.00311    | -9.44   |  |
| c.(PrevRT_Log)  | 0.04130                | 21       | 0.00337    | 12.26   |  |
| c.Trial   | -0.00002               | 0        | 0.00001    | -1.45   |  |
| <b>Distractor:</b>  |                        |          |            |         |  |
| Incongruent   | 0.01550                | 8        | 0.00313    | 4.95    |  |
| Distractor: Tone  | -0.01132               | -6       | 0.00308    | -3.68   |  |
| $R^2 = 0.3112, K = 10.946, VIF = 1.140, \text{max corr} = 0.349$  |                        |          |            |         |  |

(Congruent) should be shorter than those where they differed in articulator (Incongruent). The results of the model are shown in Table XIV. In this model as in the previous model, the control fixed effects of Intercept, SOA, Correctness of Previous Trial (Prev\_Correct)<sup>5</sup>, Previous Trial RT (PrevLogRT), and whether the response was the same as on the previous trial (PrevSame) were significant. Trial was not significant.

The Distractor effect slopes are to be compared to the baseline value of the Congruent Distractor, which is not shown separately as it is part of the Intercept value. The negative slope coefficient of the Tone Distractor shows that RTs were significantly shorter when there was a Tone Distractor than when there was a Congruent Distractor (coefficient = -0.01132,  $t = -3.68$ ). The positive slope coefficient of the Incongruent Distractor shows that RTs were significantly longer when there was an Incongruent Distractor than when there was a Congruent Distractor (coefficient = 0.01550,  $t = 4.95$ ). All collinearity values were within acceptable ranges.<sup>6</sup>

This model that included the experimental factors accounted for 31.12% of the variance in the data. Table XV shows that the model with the experimental effects

---

<sup>5</sup> Unlike in the model reported in Table IX, Prev\_correct was not centered in the model reported in Table XIV. This was due to the fact that a model with centered Prev\_correct would not converge.

<sup>6</sup> A model that included the interaction of SOA and Distractor showed that the interaction between SOA and Incongruent distractor was significant ( $t = -2.28$ ) but the interaction between SOA and Tone was not ( $t = 1.05$ ). The significant interaction reflects the fact that differences between the Congruent and Incongruent conditions at 300 ms did not show the reliable pattern of all blocks at SOAs of 100 and 200 ms. Inclusion of this interaction term did not have any material effect on the significance of the Incongruent distractor in the model. Therefore, the model including the interaction is not presented in the analysis above for comparison purposes with the model for experiment 1.

included accounted for significantly more of the variability in the response-time data than the model without them.

**Table XV. ANOVA comparison of the model with only control effects with the model that included the experimental Distractor congruency effects. The model with Distractor accounted for significantly more of the variability in the data than the model without the Distractor included, as indicated by the “Pr (> Chisq)” value.**

Models:

```

artic.Controls: Log_RT ~ c.(SOA) + Prev_correct + c.(PrevSame) +
                 c.(PrevRT_Log) + c.Trial + (1 + c.Trial | Subject) +
                 (1 | Item)
artic.Distrs:   Log_RT ~ c.(SOA) + Prev_correct + c.(PrevSame) +
                 c.(PrevRT_Log) + c.Trial + Distractor +
                 (Distractor + c.Trial | Subject) + (1 | Item)

      Df      AIC      BIC  logLik  Chisq Chi Df Pr(>Chisq)
artic.Controls 11 -11343.3 -11253.7  5682.7
artic.Distrs   20 -11411.3 -11248.4  5725.7 85.993      9 < 0.000***
```

As with the voice experiment, the significance of the effect of Congruent Distractors was tested by creating a simpler model without the subject-specific Trial and Distractor slopes. There were no qualitative changes as a result of excluding these terms, so the model not shown here due to space considerations. In this simpler model, the *t* value of the Incongruent Distractor effect was very higher at 5.22 (compared to 4.95 in the full model). Results of the Markov chain Monte Carlo sampling are shown in Table XVI.

**Table XVI. Results of Markov chain Monte Carlo sampling (with 10000 samples) for the articulator experiment. Significance determined based on the *p* value calculated by the MCMC sampling is shown in the “*pMCMC*” column. The box around the Incongruent vs. Congruent Distractor conditions show that the effect of congruency was significant with a Bonferroni-corrected alpha of 0.05/2 = 0.025.**

|                | Estimate | MCMCmean | HPD95lower | HPD95upper | pMCMC         |
|----------------|----------|----------|------------|------------|---------------|
| (Intercept)    | 6.2265   | 6.2264   | 6.1838     | 6.2658     | 0.0001        |
| c.(SOA)        | 0.0005   | 0.0005   | 0.0005     | 0.0005     | 0.0001        |
| Prev_correct   | -0.0946  | -0.0951  | -0.1108    | -0.0788    | 0.0001        |
| c.(PrevSame)   | -0.0299  | -0.0299  | -0.0364    | -0.0242    | 0.0001        |
| c.(PrevRT_Log) | 0.0462   | 0.0442   | 0.0374     | 0.0510     | 0.0001        |
| c.Trial        | 0.0000   | 0.0000   | 0.0000     | 0.0000     | 0.0001        |
| Distractor:    |          |          |            |            |               |
| Incongruent    | 0.0156   | 0.0156   | 0.0096     | 0.0214     | <b>0.0001</b> |
| Distractor:    |          |          |            |            |               |
| Tone           | -0.0114  | -0.0114  | -0.0171    | -0.0055    | 0.0001        |

The results of MCMC sampling on the simpler model of the articulator experiment showed a significant effect of the Incongruent Distractor with a *p* value calculated by MCMC sampling, with a Bonferroni-corrected alpha of 0.025, though again, see the cautionary comment at the end of section 3.2.2.2 regarding the appropriateness of this third significance test.

### 3.3.3. Discussion

The speech distractors that subjects heard in this experiment always differed in voicing from what the response was supposed to be. Congruent Distractors differed in voicing but matched in articulator, whereas Incongruent Distractors differed in both articulator and voicing. These data show that whenever there was a speech distractor, the mismatch in voicing resulted in a slowdown in RTs, since RTs were shorter when subjects heard a tone distractor than when they heard either speech distractor.

However, the penalty on RTs was less in the Congruent case than in the Incongruent case, supporting the hypothesis that RTs would be sensitive to the property of articulator and providing the first experimental evidence for a role of articulator independent of voicing in the perception-production link.

### **3.4. General Discussion**

The results across the two experiments were consistent. The control fixed effects each had a significant influence on the RTs in the task, in the directions expected, and were qualitatively the same in both experiments. The central hypotheses to be tested were also supported in both experiments: trials where subjects heard Congruent Distractors had shorter RTs than trials where subjects heard Incongruent Distractors. This effect held whether the congruency between the response and distractor was in terms of voicing or articulator. The Congruency of voicing in experiment 1 was found in data that included both tongue-tip and tongue-back responses. The Congruency of articulator in experiment 2 was found in data that included three articulators (lower lip, tongue tip, tongue body) and two manners (oral and nasal).

An effect of articulator has been noticeably absent from studies investigating the perception-production link. This is surprising given that articulator is a fundamental phonetic and phonological property and is therefore expected to play a role in the context of the increasing evidence for an intimate link between perception

and production (Gordon & Meyer, 1984). The results from the present experiment therefore provide the first experimental evidence in support of the property of articulator having a role in the interaction between perception and production. In light of the present results, it seems plausible that the reasons perceptuo-motor effects of articulator have not been found in previous studies that sought them (Galantucci et al., 2009; Gordon & Meyer, 1984; Mitterer & Ernestus, 2008) were methodological.

The absence of perceptuo-motor effects of articulator in the literature was due to such an effect being tested for (albeit with the methodological issues discussed in the introduction) but not found. The sparsity of evidence for a perceptuo-motor effect for voicing on the other hand is attributable to this being ignored by most researchers, with the notable exception of Gordon and Meyer (1984). That is, there has been a research bias toward the spatial properties of speech in the search for perceptuo-motor effects and noticeably less attention paid to the temporal properties, despite the fact that temporal properties of speech are also fundamental. Given the result for voicing of Gordon and Meyer (1984) are not unambiguously identifiable as perceptuo-motor effects, the results from the present experiments therefore demonstrate the clearest evidence to date in support of the temporal property of voicing in the perception-production link, on a par with the spatial property of articulator.

### **3.4.1. Methodological observations**

This section addresses a few methodological considerations that were discussed in the methods sections above, in light of the present data.

A few words should be said about the size of the effect in the present experiments. First, the slope coefficients for the fixed effects seem rather small. Recall that the data were log-transformed RTs, and ranged in values from about 5 to 7. Numerically small slope coefficients were therefore expected compared to response-time data in milliseconds. The significance of the effect is what is important. The ms values in Table VII and Table XIV show that the differences between conditions were roughly on the order of 4 to 8 ms. This is smaller than the differences found by Galantucci et al. (Figure 12), but again, they were significant. The meaningful observation is that the effect of congruency was significant in both experiments.<sup>7</sup> The low portion of the variance that it can account for does not mean it is not significant; it just means that the relative influence of that effect is small in comparison to, say, whether the subject made a mistake on the previous trial. This is not surprising.

Rastle and Davis (2002) concluded that intensity-based electronic voice keys, such as the one that is included with e-Prime, are unreliable in detecting the acoustic onsets of syllables, even if the onset segment is the same across syllables. The differences they found between data labeled by hand and data from electronic voice keys can in some circumstances yield statistically significant results but with opposite

---

<sup>7</sup> As Baayen (2008, p. 259) points out, there is a very low signal-to-noise ratio in response-time experiments, which is why so much data is necessary.

directions. Tyler et al. (2005) report that it is possible to build an analogue voice key that overcomes the shortcomings reported by Rastle and Davis, but this solution is not commercially available and requires the prospective researcher to custom-build the circuit. This was not viable for the present study. Rastle and Davis conclude that the most reliable method of labeling syllable onsets is by visual inspection of the acoustic waveform. This was the method used in the present experiments. Since the experiments were run using e-Prime, RTs based on the electronic voice key were recorded by default. The models reported in Table IX (experiment 1) and Table XIV (experiment 2) were re-run using the automatically recorded RTs instead of the hand-measured RTs. The results are summarized in Table XVII.

Analysis of the data from the present experiments shows that the fit of the model was worse when the models used the automatic response-time data from the e-Prime voice key. The  $R^2$  values for the models in both experiments were lower by about ten percentage points as a result of using the automatic RTs. This must be a result of the automatic RTs being more variable than the hand-measured RTs. Another consequence of using the automatic RTs was that the effect slope and its associated  $t$  value for the effect of Incongruent Distractor (vs. Congruent) were lower in experiment 1, such that the effect was not significant with the automatic RTs. In experiment 2, the effect was unchanged using the automatic RTs. It seems clear that the use of RTs calculated by electronic voice keys like the one provided by e-Prime can result in noisier data and effectively hide legitimate effects that can be seen when

RTs are calculated by visually inspecting the acoustic waveform of subjects' utterances.

**Table XVII. Hand-measured RTs vs. voice key. Comparison of the results from experiment 1 and experiment 2 depending on whether RTs were measured by hand or by electronic voice key.**

|                                   | Exp 1 hand | Exp 1 auto | Exp 2 hand | Exp 2 auto |
|-----------------------------------|------------|------------|------------|------------|
| $R^2$                             | 45.21%     | 34.12%     | 31.12%     | 20.65%     |
| Incongruent effect <i>ms</i>      | 4          | 3          | 6          | 6          |
| Incongruent effect <i>t</i> value | 2.22       | 1.90       | 4.95       | 4.95       |

One difference between this study and those reported by Galantucci et al. (2009), Kerzel and Bekkering (2000), and Gordon and Meyer (1984) is that the statistics reported here had log-transformed RTs as the dependent variable, not RTs in milliseconds. As noted in section 3.2.2.1, log-transformed RTs were used because it is an assumption of the models used here that the underlying data are normally distributed. As evident in Figure 7, the millisecond RTs are positively skewed. Let us assume that for two effects to be analyzed, one condition has shorter RTs than the other condition. The mean RT for the category with longer RTs will be numerically further away from the mean of the category with shorter RTs due to the skew in the data. This increases the probability of Type I errors. Using the log-transformed RTs mitigates this bias, and is therefore a more conservative test of the differences of the effects on RTs. This is confirmed with the data from the present experiments. When

the models reported in Table IX (experiment 1) and Table XIV (experiment 2) were re-run using the millisecond RTs instead of the log-transformed RTs, the slope for the Incongruent Distractor (vs. Congruent) effect was the same, but the associated  $t$  value of the effect was higher for both experiments (experiment 1:  $t = 2.32$  vs.  $2.22$ , experiment 2:  $t = 5.24$  vs.  $4.95$ ), which means that the analyses using millisecond RTs were less conservative.

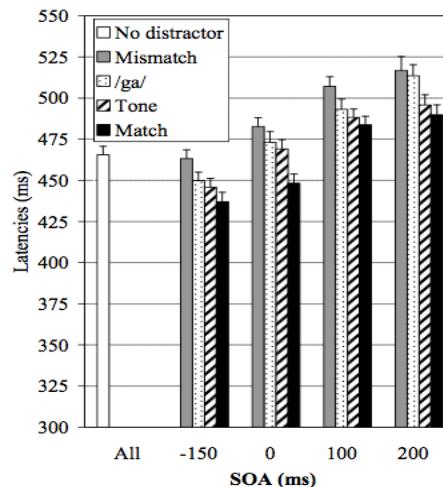
Section 3.2.2.1 motivated the inclusion of control effects in the statistical models because including them “can rule out potential confounds and increase statistical power by reducing residual noise” (Barr et al., under review). The inclusion of these effects is also advocated by, e.g., Baayen (2008). The statement of Barr et al. can be shown to be true in the present data. If the control effects are removed from the statistical models, the effect slope in ms remains the same in both experiments, but the  $t$  value for the Incongruent (vs. Congruent) effect drops in each. In experiment 1, the  $t$  value for Incongruent drops from  $2.22$  to  $1.86$ , which resulted in the effect no longer being significant. In experiment 2, the Incongruent effect remains significant, but the  $t$  value drops from  $4.95$  to  $4.82$ . Inclusion of the control effects does increase the statistical power by reducing the residual noise in the present data.

In summary, using hand-measured RTs provided more reliable data than RTs determined by an electronic voice key. Using hand-measured RTs revealed effects that were obscured by the more variable voice key RTs. Use of log-transformed RTs instead of raw millisecond values conforms better to the assumptions of the statistical

methods used and provides a more conservative but more legitimate test of the effects of the model. Including control effects is important as well to uncover small but significant experimental effects in the data.

### 3.4.2. Comparison with previous results

A review of the results from experiment 2 of Galantucci et al. (2009) is presented in Figure 12 to enlighten the discussion of the results from the two experiments reported above. The experimental designs of the two studies were similar, and further inspection will show that a direct comparison of the results of the two studies is warranted. In experiment 2 of Galantucci et al., subjects responded *ba* or *da* on all trials. “Match” distractors were identical to the response (*ba-ba* or *da-da*), while “Mismatch” distractors were the other potential response (*ba-da* or *da-ba*). At SOAs



**Figure 12. Results from experiment 2 of Galantucci, Fowler, and Goldstein (2009: modified version of Figure 2, p. 1144, reprinted with permission).**

of 100 and 200 ms, Match trials were significantly shorter than Mismatch trials, and Mismatch trials were significantly slower than Tone trials.

The inclusion of the Tone distractor and No Distractor conditions in the present experiment 2 and in Galantucci et al.'s experiment 2 allows for a direct

**Table XVIII. Comparison of distractors and responses in Galantucci, Fowler, and Goldstein (2009, “GFG”)’s experiment 2 and present experiments. The number in each cell denotes the number of features by which the response-distractor pair differed. Green denotes “Match” cases in GFG and “Congruent” cases in present experiments. Red denotes “Mismatch” cases in GFG and “Incongruent” cases in present experiments. Blank cells indicate that there was no such response-distractor pair in that block.**

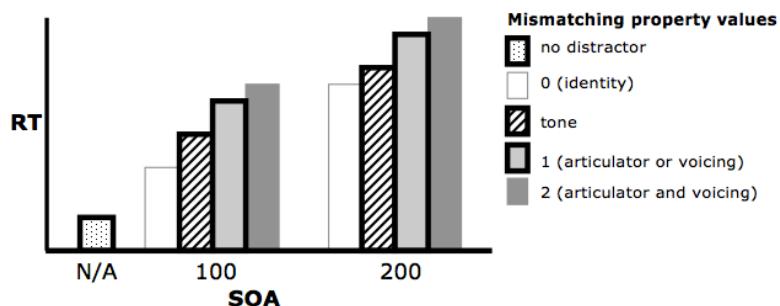
|                                |         | Distractor |           |           |           |           |             |             |
|--------------------------------|---------|------------|-----------|-----------|-----------|-----------|-------------|-------------|
|                                |         | Response   | <i>pa</i> | <i>ba</i> | <i>da</i> | <i>ga</i> | <i>tone</i> | <i>none</i> |
| GFG Exp 2                      |         | <i>ba</i>  |           | 0         | 1         | 1         | •           | •           |
|                                |         | <i>da</i>  |           | 1         | 0         | 1         | •           | •           |
| Present Voice Experiment       | Block 1 | <i>ta</i>  |           | 1         | 2         |           | •           | •           |
|                                |         | <i>da</i>  |           | 2         | 1         |           | •           | •           |
| Present Articulator Experiment | Block 2 | <i>ka</i>  |           | 1         | 2         |           | •           | •           |
|                                |         | <i>ga</i>  |           | 2         | 1         |           | •           | •           |
| Present Articulator Experiment | Block 1 | <i>pa</i>  |           | 1         | 2         |           | •           | •           |
|                                |         | <i>ta</i>  |           | 2         | 1         |           | •           | •           |
|                                | Block 2 | <i>ma</i>  |           | 1         | 2         |           | •           | •           |
|                                |         | <i>na</i>  |           | 2         | 1         |           | •           | •           |
|                                | Block 3 | <i>pa</i>  |           | 1         |           | 2         | •           | •           |
|                                |         | <i>ka</i>  |           | 2         |           | 1         | •           | •           |
|                                | Block 4 | <i>ta</i>  |           | 1         |           | 2         | •           | •           |
|                                |         | <i>ka</i>  |           | 2         |           | 1         | •           | •           |

qualitative comparison between the results in these two experiments. A comparison of the designs of the Galantucci et al. experiment and the present experiments is shown in Table XVIII. The task in both experiments was the same, and both experiments made use of 100 and 200 ms SOAs. Both experiments included the same Tone distractor<sup>8</sup> at all SOAs, and both experiments included a No Distractor condition. Note that the Match and Mismatch conditions of the Galantucci et al. experiment do not correspond to the Congruent and Incongruent conditions of the present studies. The Match trials in the Galantucci et al. study do not have a comparable case in the present experiments, because Match means that the distractor and response were the same syllable (0 in Table XVIII). It is in fact the Mismatch condition in the Galantucci et al. study that is comparable to the Congruent conditions in the present experiments since they all differed in only 1 property, although the specific properties that differed are always not the same given the differences in experimental design. The present Incongruent case has no comparable case in the Galantucci et al. study as none of their stimuli differed in more than one feature (articulator). The present experiments were designed this way to test directly and independently for effects of articulator and voicing. The only difference between the Congruent and Incongruent conditions in the present experiments was whether the response and distractor agreed or differed with regard to the property being tested in the experiment. The design of the Galantucci et al. study did not do this due to the presence of the identity condition. The contrast of interest in

---

<sup>8</sup> It was in fact the same audio file used by Galantucci et al. (2009), which was kindly provided by Bruno Galantucci to the author for use in the present experiments.

the Galantucci et al. experiment was between the identity condition (the “Match” condition in their terminology) vs. a condition where the response and distractor differed on the property of articulator (“Mismatch”). Therefore, the Galantucci et al. Mismatch condition was effectively equivalent to the Congruent condition in the present experiment 2.



**Figure 13. Schematic, qualitative comparison of distractor conditions based on Experiment 2 of Galantucci, Fowler, and Goldstein (2009) and the results from the present experiment 2. Distractors differed from a required responses on either 0 parameters (the identity condition), 1 parameter, or 2 parameters. Conditions that were included in both experiments (no distractor, tone, and 1 mismatching parameter) are indicated by bars with bold borders. The present experiment compared conditions where distractors differed from required responses on 2 parameters (the Incongruent condition) with the condition where distractors differed from required responses on 1 parameter (the Congruent condition). The Galantucci et al. experiment compared conditions where distractors differed from required responses on 1 parameter (the Mismatch condition) with the condition where distractors differed from required responses on 0 parameters (the Match condition). The Galantucci et al. experiment included the identity condition, which the present experiments did not, and the present experiments included the condition with 2 mismatches, which the Galantucci et al. experiment did not. Both experiments included the same tone distractor condition, and a condition with no distractor. The condition where distractors differed from required responses on 2 parameters was slower than the condition where distractors differed only on 1 parameter (results from present experiment 2). The condition where distractors differed from required responses on 1 parameter was slower than the tone condition (both experiments). The condition where distractors differed from required responses on 0 parameters was faster than the condition where distractors differed only on 1 parameter (Galantucci et al. experiment), and faster than the tone distractor condition.**

A schematized comparison of the results of the present experiment 2 and experiment 2 of Galantucci et al. is shown in Figure 13. The results for the condition in each experiment where the responses and distractors differed in articulator (Mismatch and Congruent) were qualitatively the same in both experiments: at 100 and 200 ms SOAs, both the Congruent condition in the present experiment 2 and the Mismatch condition in the Galantucci et al. experiment were slower than the Tone condition. By transitivity, the Incongruent condition of the present experiment 2 can be compared with the identity (Match) condition of the Galantucci et al. experiment because each of these conditions was compared within their experiment to the tone conditions and the condition with 1 mismatching property. The Congruent and Incongruent conditions in the two experiments presented here can be equated because they differ only in which single property is manipulated in the Congruent conditions. Drawing on data from both experiments, we can therefore confidently state that RTs showed the following pattern (using the terminology of the conditions from the present experiment): no distractor < identity < tone < Congruent < Incongruent. Response times were increasingly longer with an increasing number of mismatching properties between the response and the distractor.

### **3.5. Conclusion**

In summary, identity between a distractor and a response has been shown to have facilitative perceptuo-motor effects on response times in other studies, and identity may have a privileged status in facilitating speech production. The main purpose of the experiments reported in this study was to determine whether specific temporal and spatial properties of speech, voicing and articulator, respectively, are involved in the interaction between speech perception and speech production, as there has been a noticeable lack of such results in the literature. Two experiments using a response-distractor task provide the first clear experimental evidence that, in normal speaker-hearers, both of these properties do exert an independent influence on speech production. Response times on trials when subjects heard a spoken distractor syllable that differed in both voicing and articulator from the syllable they produced were longer than on trials where the distractor differed only in articulator (experiment 1) or voicing (experiment 2). This result was shown to be consistent with previous studies. In light of the present results, it does not seem necessary to appeal to a special status for identity in the perception-production link. A minimal requirement of any model of the perception-production link, then, is that the model accommodate perception-production interaction at the level of the properties of articulator and voicing. Such a model is developed in chapter 6.

## **CHAPTER 4: MOTIVATING EXCITATION AND INHIBITION**

### **4.1. Introduction**

The goal of this chapter is to demonstrate that the experimental results from Chapter 3 and other results in the literature require that any computational model of the interaction between speech perception and production include the principles of both excitation and inhibition. The experimental evidence motivating a role for both excitation and inhibition is presented in section 4.2. Section 4.3 addresses some potential counter-arguments to the motivations and assumptions made in section 4.2, ultimately concluding that seemingly contradictory results and models in the literature do not undermine the assumptions made in section 4.2. Section 4.4 concludes.

### **4.2. Experimental evidence**

A comparison of the experimental results from Chapter 3 (henceforth, “Experiment 1” and “Experiment 2”) along with studies by Kerzel and Bekkering (2000) and Galantucci, Fowler, and Goldstein (2009) reveals a clear pattern of perceptuo-motor effects on RTs. Close examination of the perceptuo-motor effects in the light of non-linguistic influences on RTs found across all of these experiments will show that the two computational principles of excitation and inhibition are both required to explain the experimental results.

#### **4.2.1. Linguistic and non-linguistic influences on RTs**

The results from both Experiments 1 and 2 show two consistent non-linguistic influences on RTs: the presence of any distractor vs. no distractor, and SOA. Comparing the No Distractor condition in Experiment 1 (Figure 5 from Chapter 3) and Experiment 2 (Figure 10 from Chapter 3) to all conditions where there was a distractor of any kind, it is clear that the presence of a distractor increased RTs, regardless of whether the distractor was a speech syllable or a tone.<sup>1</sup> This influence of distractor vs. no distractor was also found by Galantucci et al. (2009) for SOAs shared between their experiment 2 and both Experiments 1 and 2 (100 and 200 ms). In addition, a significant influence of SOA was found in both Experiment 1 and 2 when a distractor of any kind was present, with RTs increasing monotonically as SOAs increased. This replicates the SOA findings of both Galantucci et al. (2009)<sup>2</sup> and Kerzel and Bekkering (2000). It is clear that the influences of presence vs. absence of a distractor and SOA were qualitatively the same in both of Experiments 1 and 2 and in experiments by other researchers using the same or very similar experimental task (see

---

<sup>1</sup> Mixed-effect linear models were created of the log-transformed RT data from Experiments 1 and 2 (one model for each experiment) of Chapter 3 to test whether this difference was significant. All trials with a RT greater than zero and less than 1500 ms were included. A fixed effect of SOA with four categorical levels was included: “NA” was the No Distractor condition, and the 100, 200, and 300 levels represented all trials at that SOA across all Distractor conditions. Distractor type was not included in the model. RTs for the No Distractor (i.e., SOA “NA”) condition were significantly shorter than for each of the other three SOAs, in both experiments (all  $t$  values  $> 3.5$  and all MCMC-calculated  $p$  values  $< .001$ ).

<sup>2</sup> RTs when the distractors had 0- and -150-ms SOAs were faster than the No Distractor condition in the study by Galantucci et al. (2009). Their interpretation of this result is that these distractors alert subjects that the presentation of the stimulus is imminent, which primed subjects to respond more quickly. This is in contrast to the conditions when the distractor followed the visual cue, where the distractor consistently slowed the subjects down during the process of preparing the response on a given trial.

Figure 13 from Chapter 3), indicating that these non-linguistic influences presumably arise from some other cognitive process (or processes) involved in this task that do not involve the perception-production link. For simplicity's sake, this set of processes will be referred to collectively as Distraction Processing. The RT slowdown of the Tone condition at various SOAs compared to the No Distractor condition therefore can be treated as a baseline RT reference indicating the influence of Distraction Processing, but not reflecting any influence of the process that generates perceptuo-motor effects. RTs in the Congruent/Incongruent conditions in Experiments 1 and 2 then reflect the influence of Distraction Processing combined with the influence of the process that generates perceptuo-motor effects, which is proposed in Chapter 6.

#### **4.2.2. Excitation and inhibition**

To recap the discussion in Chapter 3, the pattern of RTs is:

no distractor < identity < tone < congruent distractor < incongruent distractor.

This sub-section examines the results from Experiment 2 and experiments of others to show that the computational principles of both excitation and inhibition are required. The evidence for the presence of excitation and inhibition can be seen by considering the interaction between Distraction Processing and the process that accounts for the perceptuo-motor effects, specifically by comparing the various linguistic conditions from Experiment 2 and from the Galantucci et al. (2009) experiment with the No Distractor and Tone conditions. The crucial facts to be accounted for are that RTs in

the Identity condition were shorter than the Tone condition, while they were longer in the Mismatch/Congruent condition than in the Tone condition, and longer still in the Incongruent condition.

If the process that produces perceptuo-motor effects introduces only excitation, the facilitation observed in the Identity condition of the Galantucci et al. (2009) experiment can be explained, but the results from Experiment 2 cannot. One could reasonably propose an excitation-only process that introduces perceptuo-motor facilitation only for the Identity condition. If this is the case, then all other distractor conditions should be slower than the Identity condition, but no differentiation between distractor types is predicted. The results from Experiment 2 are not consistent with this prediction, since RTs increased with each mismatching parameter between the required response and the distractor, i.e., the RTs were longer for the Incongruent condition than for the Congruent condition. Another possible excitation-only process could be reasonably proposed in which increasingly more excitation is introduced with each matching parameter, compared to some baseline. Such a process could account for the cross-experiment results shown in Figure 13 of Chapter 3, where RTs increased monotonically with the number of mismatching parameters (0 vs. 1 vs. 2). However, an account based on such an excitation-only process predicts that the Tone condition should be the slowest condition, since it benefits from no perceptuo-motor facilitation. This is not what was found in Experiments 1 and 2. RTs in the Tone condition were shorter than in both the Congruent and Incongruent conditions, in which the distractor

differed from the required response on 1 and 2 parameters, respectively. Therefore, any account where the process generating perceptuo-motor effects through excitation only seems implausible.

An alternative that can also ultimately be excluded is that the process introducing perceptuo-motor effects relies exclusively on inhibition, where the amount of inhibition is proportional to the number of mismatching properties. An account based on inhibition only would be consistent with the results from Experiments 1 and 2. The Congruent condition was slower than the Tone condition, and the Incongruent condition was slower than the Congruent condition, treating the Tone condition as the baseline reference where there were no influences from the process that introduces perceptuo-motor effects. However, an inhibition-only process could not result in the facilitation found by Galantucci et al. (2009) where the Identity condition had shorter RTs than the Tone condition.

It is clear then that any account of the perceptuo-motor effects found in these studies must include the principles of both excitation and inhibition. The experimental results indicate that excitation and inhibition are introduced by interactions between inputs at the level of phonological features. Matching features introduce excitation, and mismatching features introduce inhibition. Since these RT modulations in the experiments reflect perceptuo-motor effects (see section 2.3 in Chapter 2), the excitation and inhibition introduced by the distractors should be influencing the

process of phonological planning. The details of this interaction are spelled out in the model presented in Chapter 6.

#### **4.3. Do “common codes” excite or inhibit?**

The assumption in the above interpretation of the experimental data presented in Chapter 3 is that shorter RTs represent the simultaneous activation of some representation—or “code”, to use the term of Viviani (2002)—by processes of both production and perception. When perception and production activate the same representations, the activation level of the representations involved in the production system breaches threshold sooner than it would otherwise (see section 2.3.2 of Chapter 2). For example, in the experiments in Chapter 3, when a perceived distractor has the same voicing feature value as an utterance that is being planned, the activation level of the voicing feature of a response rises quicker than it would in absence of the reinforcing input that is the result of perception (all other things being equal). This results in shorter RTs. This assumption is commonplace in the literature on the interaction between perception and production (e.g., Fowler, Brown, Sabadini, & Weihing, 2003; Galantucci et al., 2009; Gordon & Meyer, 1984; Kerzel & Bekkering, 2000; Mitterer & Ernestus, 2008, among many others).

There are experimental results, however, that suggest that use of common codes may result in longer RTs, and theories that assert precisely that. If this were the case, the basic interpretation of the results from Chapter 3 presented above would be

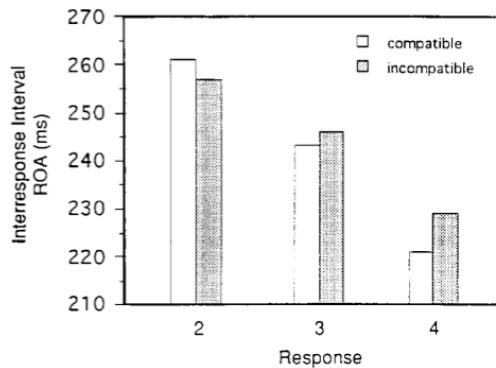
called seriously into question. Careful consideration of the data and theories of these studies shows that they do not undermine the principles that are motivated in this chapter.

#### **4.3.1. Accessing a common code**

Prinz (1997) maintains that longer response latencies, not shorter ones, provide evidence for links between perception and production. Prinz's claim is that a code being used by one process (e.g., perception) is unavailable to a different process (e.g., production) until the first process is finished with that code. This claim will be referred to here as the theory of "code inaccessibility". Code inaccessibility predicts that latencies should be longer when two processes attempt to access the same code at the same time. Prinz (1997, pp. 147–149) reports a study by Müsseler (1995) as evidence in support of this theory. In the experiment, subjects had to make a series of various orderings of five left and right key presses on each trial. At the start of each trial, the subject saw a sequence of four left and right arrow combinations on a computer screen. The arrows indicated the sequence of the first four key presses. On any given trial, an instruction for a fifth key press appeared after the subject pressed the first key of the four-arrow sequence for that trial. The measurement of interest was the latency between the first and second key presses. Code inaccessibility predicts that this latency should be longer when the direction of the fifth arrow cue is the same as that of the second key press compared to when it points in the other direction because the

subject's production planning processes is using one of the two codes (RIGHT or LEFT) immediately after the first key press to prepare for the second key press. At the same time, subjects see another arrow which requires them to append the fifth key press direction to the queue they are in the middle of executing. Adding this direction to the queue requires accessing either the RIGHT or LEFT code. If the arrow directions are the same, the code for producing the second key press is not available to that production process until it is "freed up" from the process that is adding it to the queue. This should result in longer latencies for this key press. On the other hand, code inaccessibility predicts that if the arrow directions are different, there should be no increase in latencies because there is no competition for codes.

Figure 14 shows the results of the Müsseler experiment. The latencies between the first trial and second trial (labeled "2") were longer when the instruction arrow direction was the same as the key press (compatible) compared to when the arrows were different (incompatible). Prinz interprets the result for the second key press as support for code inaccessibility. However, based on the results of the Müsseler experiment alone there is reason to question both some of the assumptions underlying code inaccessibility, and whether it applies to the link between perception and production planning.



**Figure 14. Results of the interference task with arrow presses of Müsseler (1995), as reported in Prinz (1997, p. 148, his Figure 8a). Reprinted with permission.**

Note that the pattern for latencies between trials 2 & 3 and 3 & 4 (labeled “3” and “4”, respectively, in Figure 14) was the opposite from the latencies between the first and second key presses: latencies on the third and fourth key presses were shorter when the arrow key pressed was the same as the instruction arrow (compatible) than when it was different (incompatible). The differing pattern of results across the different key presses suggests that there is more than one process influencing latencies, and that codes do not interact with all processes uniformly. Let us posit that there are four processes involved in this task. The first is the process that perceives the cue arrows presented in the task. The second is a queue building process that makes the list of key press directions to be executed. The third is a production planning process that loops through the queue serially and sends each direction specification of LEFT or RIGHT to a fourth implementation process that effects the key presses of the appropriate arrow key.

Just before the second key press, the first three of these processes are presumably involved. Subjects must perceive the cue arrow (process 1), incorporate into the queue of key presses what direction the fifth key press needs to be (process 2), and plan the direction of the second key press from the queue (process 3). If we assume that the fifth key press is incorporated into the queue by the time the second key press is executed, then the queue building process should not be involved in the third or fourth key presses. The queue building process and its interaction with either or both of the other two processes may be the cause of the longer latencies on the second key press, and possibly due to code inaccessibility. This influence of code inaccessibility on latencies should and does then disappear on subsequent key presses.

However, the fact that latencies in the compatible condition were shorter than in the incompatible condition for the third and fourth key presses is problematic in and of itself for code inaccessibility, which does not predict any difference in latencies if there is no competition for a particular code. Prinz states that the third and fourth key presses show “a normal compatibility effect” (Prinz, 1997, p. 148), by which he presumably means an effect where common codes used by both perception and production yield facilitation. These “normal compatibility effects” must be the result of the interaction between some set of processes other than the set of processes in effect on the second key press. From the four processes above, the production planning process (3) must be engaged, and as just discussed, the queue-building process (2) is not. The “reversed” latency effects on key presses 3 and 4 could be

explained by the interaction between the cue perception process (1) and the ongoing production planning (3), if use of a common code in the interaction between these processes has a facilitatory effect.

This explanation is problematic for the code inaccessibility theory, which posits that the use of a code by any process makes it inaccessible to any other, since this does not seem to be the case when the perception process and the production planning process make use of the same code. The code inaccessibility theory therefore does not seem to be adequately supported by the data from Müsseler (1995). Even if some type of code accessibility is the cause of the compatibility effect observed for the second key press, this seems to be a result of involving the queue building process. However, when the perception and production planning processes share a code, the effect is the opposite, resulting in shorter response latencies. This is consistent with the standard view of facilitatory effects of stimulus compatibility on response latencies. Prinz's theory of code inaccessibility therefore does not pose a serious problem to the view adopted here that shorter response latencies in the experiments reported in Chapter 3 reflect facilitation introduced by the interaction of perception and production.

#### **4.3.2. Repeated phonological planning**

Some experimental studies have found inhibitory effects on speech production arising from phonemic or featural similarity across consecutive utterances. The

findings from both of these studies are presented below. A brief discussion follows showing that the nature of the processes involved in these tasks are very different from those involved in the experimental task used in Chapter 3, and ultimately do not pose a problem for the principles motivated above for explaining the response-distractor task.

In an experiment reported by Sevald and Dell (1994), subjects had to repeat sequences of four real English CVC words as many times as they could in a fixed amount of time. The syllables in the sequences were either identical (e.g., *pick-pick-pick-pick*), or varied by initial consonant (e.g., *pick-tick-tick-pick*), vowel (e.g., *pick-puck-puck-pick*), final consonant (e.g., *pick-pin-pin-pick*), CV (e.g., *pick-tuck-tuck-pick*), and VC (e.g., *pick-pun-pun-pick*). Subjects' speaking rate was fastest, i.e., they produced the most repetitions of the sequences, when the sequence was identical or repeated the final consonant (e.g., *pick-tuck-tick-puck*). Their speaking rate was significantly slower when the initial consonant was repeated (e.g., *pick-pun-pin-puck*), and when the CV was repeated (e.g., *pick-pin-pin-pick*). The authors attribute these slow downs in speaking rate to competition at the phoneme level, as defined by a phonological competition model of O'Seaghda, Dell, Peterson, and Juliano (1992). The relevant aspect of that model for present purposes is that it predicts when two words are spoken in succession, RTs should be longer when certain phonemes are repeated but the words are not identical, due to competition introduced by the two words sharing a phoneme node. Use of a common phoneme code across utterances is therefore inhibitory, not facilitative.

Rogers and Storkel (1998) report a series of experiments where subjects read series of real English monosyllabic words that appeared on a computer screen one at a time. After repeating a prime word a few times in a row, a target word would appear whose first phoneme varied with regard to how many features it shared with the prime word. All primes and targets ended in /-ʌg/. RTs for prime-target pairs whose initial phone had no features in common (e.g., *chug-bug*) served as a baseline. The four experimental conditions were prime-target pairs that shared: manner only (e.g., *tug-bug*), voicing only (e.g., *jug-bug*), place and manner (e.g., *pug-bug*), and voicing and manner (e.g., *dug-bug*). They ran four experiments where they varied the length of the inter-stimulus interval and number of words in the set of responses. Across the four experiments, no conditions ever had shorter RTs than the baseline. Shared manner most often resulted in significantly longer RTs, specifically, in the shared-manner and shared-manner-and-voicing conditions (e.g., *tug-bug* and *dug-bug*), but not in the shared-place-and-manner condition (e.g., *tug-bug*). No other condition resulted in reliable differences in RT. The authors interpret their result as evidence for planning of phonetic-level features, whose specific values are independently inhibited after being sent to motor implementation. When two words in succession need to make use of the same feature code, RTs are slower because the inhibited shared feature code has to overcome this inhibition to reach threshold level for the second utterance. Thus, shared codes introduce inhibition, not excitation.

While the experimental tasks used in both the Sevald and Dell (1994) and Rogers and Storkel (1998) studies are different in many ways, what they have in common is a focus on how activation levels of various codes get reset after an utterance has been made. Such a mechanism is common in models of speech production (e.g., Dell & O'Seaghda, 1992; MacKay, 1987; Shattuck-Hufnagel, 1979) to ensure that a code that has just been activated does not get re-selected by a subsequent utterance (or subsequent part of the same utterance). The processes involved in the above studies both crucially rely on the nature of the immediately planned or produced utterance. The effects of congruency found in the response-distractor task can be explained without any crucial reference to the nature of a previously planned utterance (see section 6.3.5 of Chapter 6 for more details). Therefore, the findings of the authors above do not pose a problem for the assumptions here that inhibition and excitation interact within the planning of a single utterance as motivated in section 4.2. Indeed, the common view that activation levels get actively lowered after selection predicts that effects across utterances should be different from effects within utterance planning.

#### **4.3.3. Revised phonological planning**

Another set of experiments has found that when subjects are forced to change their phonological plan before producing an utterance, RTs are longer if the original and revised utterance have target phonemes that are more similar than when they are

less similar. RTs are shorter only in the case of repeated phonemes. However, the model proposed by the authors of these studies to account for the effects they found operates at a level of representation that is different from the level involved in the model in Chapter 6. Disparities in the effects found in these studies and the experiments from Chapter 3 suggest different processes operating at different levels of representation, with neither set of results being necessarily problematic for the other.

Meyer and Gordon (1985) report experimental results which provide evidence that preparing to produce one utterance can slow down the production of an alternate utterance whose final consonant has the same articulator or voicing of the first utterance. The particular task they used was a response-priming procedure, where subjects saw two syllables written in English orthography, e.g., “UB UT”. Subjects then heard one of two tones. A high-pitch tone indicated that subjects should produce a ‘primary response’, which was either both syllables as presented on the screen (e.g., *ub-ut*), or just the first syllable of the pair (e.g., *ub*), depending on the experiment. A low-pitch tone indicated that they should produce a ‘secondary response’, which was either the syllables in reverse order from how they were presented (i.e., *ut-ub*), or the second syllable of the pair (e.g., *ut*), again, depending on the experiment. Pairs of responses were divided into four conditions, one where the final consonants either differed in both voicing and articulator (e.g., *ut-ub*, i.e., the nothing-shared condition), one where they matched in voicing and differed in articulator (e.g., *ub-ud*, i.e., matching articulator), one where they matched in articulator and differed in voicing

(e.g., *ub-up*, i.e., matching voicing), and one where they matched in articulator and voicing (e.g., *ub-ub*, i.e., identity). Meyer and Gordon assumed that on all trials subjects highly prepared to make the primary response, and that subjects had to quickly re-plan their responses on trials where they were required to make the secondary response. RTs on secondary-response trials were the dependent measure, to see what the effect of the different conditions would be on the time it would take subjects to re-plan their response. The nothing-shared condition served as a baseline. They found that RTs for secondary responses in the identity condition were significantly shorter than the baseline, but RTs for secondary responses in both the matching articulator and matching voicing conditions were significantly longer than the baseline condition. They “infer[ed] that phonetic features play a significant role in central preparation and articulatory motor programming” (Meyer & Gordon, 1985, p. 20), based on the longer RTs—rather than shorter—when the primary and secondary responses shared articulator or voicing.

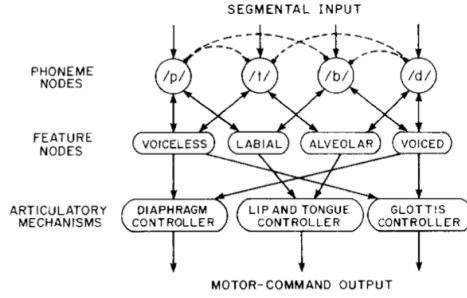
Meyer and Gordon account for these RT differences with an interactive-activation model of syllable production, shown in Figure 15. In their model, preparation of a syllable involves activating the required phoneme nodes. An activated phoneme node has two effects on the model. First, it activates all of the feature nodes associated with that phoneme. These activation-spreading links are indicated by the solid arrows in Figure 15. Feature nodes in turn activate the necessary articulatory mechanisms. Second, the model stipulates that there is an inhibitory link between two

phoneme nodes for each feature node that the two phonemes share. These links are indicated by the dotted arrows in Figure 15. These inhibitory links provide the key to accounting for the experimental results on trials where the subject has to produce the secondary response. By the authors' account, on every trial, the subject highly prepares for the primary response, e.g., *up*. As a result, the activation level of the /p/ phoneme node is high, as are the associated feature nodes of voiceless and labial. However, on secondary-response trials the subject has to re-plan the utterance, and the starting point of the activation levels for each node is where it was as a result of the subject having prepared the primary response. Therefore, the nothing-shared condition (here *ud*) is the baseline, since the activation level of the /d/ phoneme node is unchanged from its resting position. This is because /p/ does not inhibit /d/, as they do not share any feature nodes.<sup>3</sup> In the matching articulator or voicing case (here, secondary responses of *ut* and *ub*, respectively), activating the /p/ phoneme node inhibits the /t/ phoneme node because they share the voiceless feature node. /p/ also inhibits the /b/ phoneme node because they share the labial articulator node. RTs are longer than the baseline because the activation levels have to rise further to overcome the lower starting level resulting from the inhibition from /p/. In the identity condition,

---

<sup>3</sup> The phonemes /p/ and /d/ do share many features, e.g., [-continuant], [+consonantal], etc. (Chomsky & Halle, 1968). Meyer and Gordon restrict the discussion of their model to articulator and voicing features only, and are unclear as to how the inhibition works when a complete set of feature nodes is implemented. It seems reasonable to assume in their model that there is an inhibitory connection between two phoneme nodes for each feature node they share, and that the effects of the inhibitory connections are therefore roughly additive. Thus, activating /p/ inhibits /t/ and /b/ *relatively* more than it inhibits /d/.

RTs are shorter than the baseline because the activation level of the required phoneme node /p/ is already elevated as a result of the subject having prepared for it.



**Figure 15. Interaction-activation model of syllable production from Meyer and Gordon (1985, p. 20, Figure 2), showing a trial where the primary response was *ut-ub* and the secondary response was *ub-ut*. Reprinted with permission.**

Neither the experimental results nor the model from the Meyer and Gordon (1985) study necessarily poses a problem for interpreting shorter RTs in the Chapter 3 as reflecting role for phonological features in the link between perception and production. The longer RTs for shared voicing and articulator are accounted for in their model entirely by the inhibitory links between the phoneme nodes. Their model would still account for their results without the featural-node and articulatory-movement node tiers, assuming no further complicated interactions among nodes below the phoneme tier, which are not proposed (Figure 15). The feature-node tier plays no role whatsoever in deriving the RT effects they found. The model of Meyer and Gordon (1985) should therefore have no influence on an account of the results from Chapter 3 that involves only processes at the level of phonological features. The results from the experiments in Chapter 3 cannot arise from the phoneme-level inhibition in the model of Meyer and Gordon (1985). In the experiments from Chapter

3, trials with Incongruent distractors mismatched in both voicing and articulator (e.g., *ta-ba*) and were slower than trials with Congruent distractors, which mismatched only on one of voicing (e.g., *ta-da*) or articulator (e.g., *ta-pa*). According to the Meyer and Gordon model, if effects of congruency in the response-distractor task were due to interactions at the phoneme level, RTs in the Incongruent condition should have been shorter than in the Congruent condition because the perceived distractor phoneme would not inhibit the phoneme being planned since they share no features. This is the opposite prediction from what was found in the experiments in Chapter 3.

Yaniv, Meyer, Gordon, Huff, and Sevald (1990) report a series of experiments that use a similar response-priming task to the one used by Meyer and Gordon (1985). Subjects had to prepare spoken CVC responses, with different prompts indicating whether they should produce a primary or secondary response. Primary-secondary pairs were manipulated such that the vowels were either the same (e.g., *peak-beat*), similar (e.g., *peak-putt*), or dissimilar (e.g., *peak-pot*). The dependent variable was the RT of the secondary responses. RTs were significantly different between three conditions. RTs were longer when the vowels were similar (i.e., /i/-/ɪ/ or /ʌ/-/a/) than when they were dissimilar (e.g., /i/-/a/). That is, RTs were longer when the vowel of the primary and secondary responses differed in one feature ([±tense]) only than when they differed in 2 ([±high] and [±back]) or 3 ([±high], [±back], and [±tense]). The other significant difference was that RTs on secondary responses where the vowels

were identical were shorter than in the dissimilar condition. Yaniv et al. (1990) suggest that their results are consistent with the model with phoneme-node inhibition proposed by Meyer and Gordon (1985). According to their account, when the subject has to change his or her phonological plan to produce a different vowel, phoneme nodes that share many features with but are not identical to the vowel of the primary utterance are inhibited by planning the vowel of the primary response. Re-planning the secondary response requires overcoming the lower activation of the required vowel phoneme node. In the identity condition, no such inhibited vowel-phoneme node has to be re-programmed, so RTs are shorter.

Assuming that the effects observed by Yaniv et al. (1990) are accounted for by the model of Meyer and Gordon (1985), the inhibition they found for similar vowels does not pose a problem for the motivation of excitation and inhibition in the experimental results from Chapter 3 as laid out in section 4.2, for the same reasons as just outlined above with regard to the Meyer and Gordon (1985) results. In the task in Chapter 3, subjects simply plan and produce an utterance. As argued in section 4.2, RTs are modulated by excitation and inhibition introduced to the planning process at the level of phonological features by properties of a perceived distractor. If an account of the congruency effects reported in Chapter 3 does not refer to or rely on the state of phonemes nodes during the planning process, then the effects predicted by the model of Meyer and Gordon (1985) are irrelevant, especially since the task used in the experiments from Chapter 3 does not involve any re-programming of phonological

features. The model developed in Chapter 6 will not refer to or rely on phoneme-level nodes, so the findings of Meyer and Gordon (1985) and Yaniv et al. (1990) do not pose a problem in accounting for the effects observed in the task used in Chapter 3. The model and experimental results of Meyer and Gordon (1985) therefore support locating the interaction between perception and production in the task used in the experiments from Chapter 3 at a level other than phoneme activation, i.e., at the level of phonological features.

#### **4.4. Conclusion**

This chapter establishes that an adequate account of the experimental results in Chapter 3 must include the computational principles of excitation and inhibition, and a role for representations at the level of phonological features. Mismatching features between a perceived distractor and an utterance being planned introduce inhibition, which matching features introduce excitation. Close examination of studies that appear to call these conclusions into question showed that the theoretical assumption that use of common representations should result in facilitation is sound, and that examples of phonological similarity introducing inhibition are the result of processes that are not relevant to the response-distractor task and that operate on different levels of representation. The next chapter introduces a computational framework that will be used in the following chapter to model the experimental results from Chapter 3. This framework has at its core the principles of lateral inhibition and local excitation.

## **CHAPTER 5: DYNAMIC FIELD THEORY AND A PHONEME**

### **CLASSIFICATION TASK**

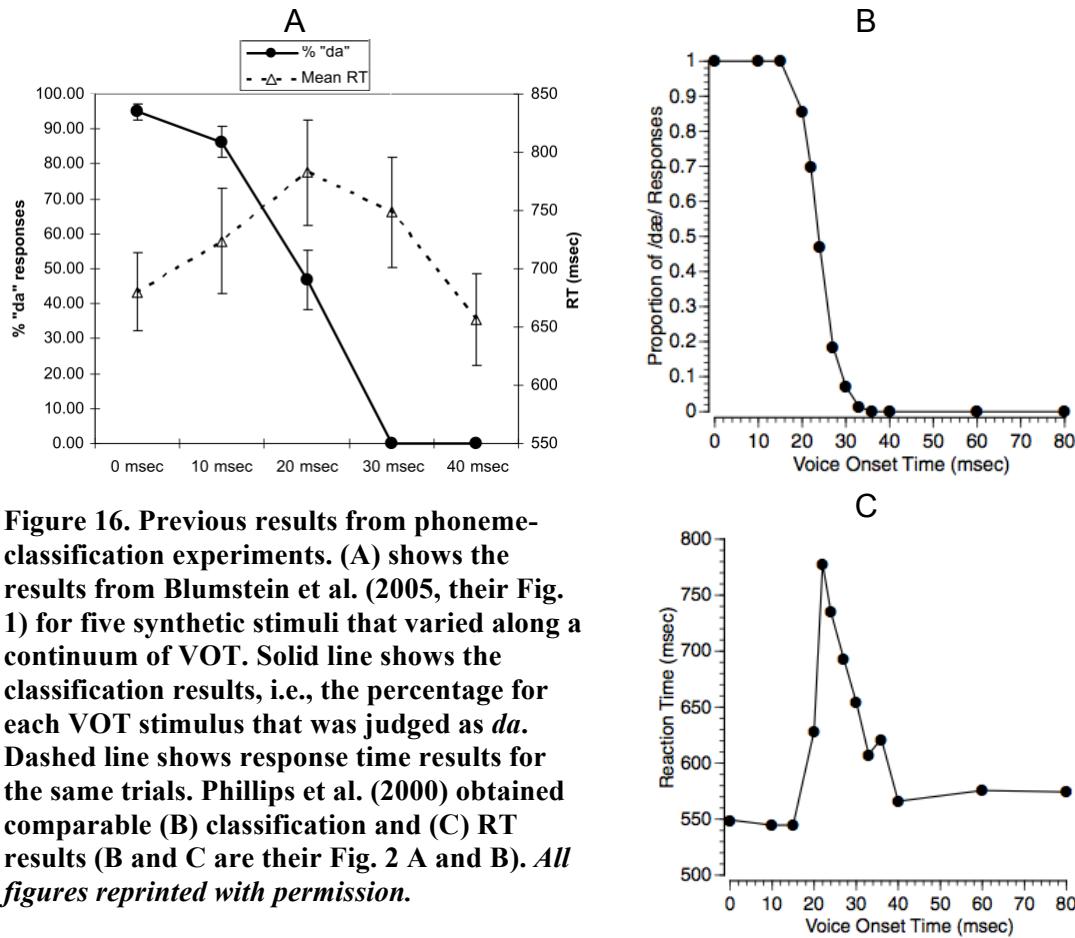
#### **5.1. Introduction**

This chapter introduces the computational framework that will be used in Chapter 6 to model the dynamics of phonological planning in the response-distractor task used in the experiments in Chapter 3. The framework is introduced by modeling a phoneme classification task. Section 5.2 explains the experimental task that was modeled. Section 5.3 then introduces the model and the computational framework used to create it, and presents the results from a simulation of the phoneme classification experiment using the model. In addition to explaining the principles and details of the computational framework, this section will show that the model can use a single set of computational mechanisms to account for both categorical classification data and gradient response time data in the same task. Section 5.4 discusses the model and the results of the simulation. Section 5.5 concludes.

#### **5.2. Phoneme classification**

Many studies have shown that subjects demonstrate a categorical perception of phonemes when presented with stimuli that vary across some phonetic continuum. Stimuli on one side of some point along the continuum are perceived equally as belonging to one category, while stimuli on the other side are perceived as belonging

to a different category. Stimuli at or very near to this “phoneme boundary” are perceived ambiguously. A continuum of Voice Onset Time (“VOT”, Lisker & Abramson, 1964) has been experimentally manipulated to demonstrate the categorical perception of phonemes differing in voicing (e.g., Blumstein, Myers, & Rissman, 2005; Eimas & Corbit, 1973; Phillips et al., 2000, among many others). Gradient



manipulation of F2 formant transitions into a vowel has been shown to have similar categorical effects on the perception of the place of articulation of the onset consonant

(e.g., Dehaene-Lambertz, 1997; Liberman, Harris, Hoffman, & Griffith, 1957; Pisoni & Tash, 1974; Werker & Lalonde, 1988, again, among many others).

The study by Blumstein et al. (2005) is typical of this type of result. They presented subjects with five different synthesized CV auditory stimuli that had VOT values of 0, 10, 20, 30, or 40 ms. The formant transitions were synthesized such that the initial consonant was always identifiable as a coronal stop, i.e., *ta* or *da*. Subjects had to indicate by pressing a button whether they thought the stimulus they heard was *ta* or *da*. Their results are shown in Figure 16A. There were three groups of classification responses that were significantly different from each other: 0- and 10-ms stimuli (unambiguously *da*), 20-ms stimuli (ambiguous), 30- and 40-ms stimuli (unambiguously *ta*). The differences within those groups were not significantly different from each other, see the solid line in Figure 16A. The results showed that stimuli were classified categorically, except for the stimulus whose VOT fell in a narrow band between the two categories (20 ms), which was classified roughly half the time as *da* and half as *ta*.

Blumstein et al. (2005) recorded not only the classifications, but also the response times (RT), defined as the time between the onset of the audio stimulus and the button press. These results are shown by the dot-dashed line in Figure 16A. The RT results also show three groups of RTs that were significantly different from each other, though not the same three groups as with the classification data: 0- and 40-msec (“good” exemplars of their respective phonemes), 10- and 30-msec (“marginal”

exemplars), and 20-msec (“ambiguous” exemplars). The good stimuli were classified most quickly, the ambiguous stimuli most slowly, and the marginal stimuli in between. In summary, the classifications were sensitive to the categorical VOT boundary, whereas the RTs were sensitive to the category “goodness” of the stimuli. Phillips et al. (2000) found effectively the same qualitative results for both classifications (Figure 16B) and RTs (Figure 16C) using the same experimental design, for stimuli on a *tae-dæ* continuum.

### **5.3. Model and simulations**

This section has two main goals. The first is to introduce and explain the computational framework that is used to build both this model and the model that is presented in Chapter 6 to account for the experimental results in the response-distractor task from Chapter 3. The second is to provide a single computational model of the classification task that can account for both the categorical classification results and the gradient RT results. Section 5.3.1 provides background on the specific formal computational framework used in the model. Section 5.3.2 presents and explains the model of the phoneme classification task. The results of simulations of the phoneme classification task are reported and discussed in section 5.3.3.

#### **5.3.1. Dynamic Field Theory**

The model of the phoneme-classification task was implemented using Dynamic Field Theory (“DFT”), a theoretical mathematical framework for modeling

movement planning developed by Schöner and colleagues (Erlhagen & Schöner, 2002; Kopecz & Schöner, 1995; Thelen, Schöner, Scheier, & Smith, 2001, and many others).

In DFT,

“[m]ovement parameters are represented by activation fields, distributions of activation defined over metric spaces. The fields evolve under the influence of various sources of localized input, representing information about upcoming movements. Localized patterns of activation self-stabilize through cooperative and competitive interactions within the fields” (Erlhagen & Schöner, 2002: p. 545).

The specifics of the activation fields and the interactions of inputs to them will be illustrated in detail in this chapter using the model of the phoneme classification task.

Chapter 4 showed that the experimental data presented in Chapter 3, in combination with experimental results of others using the same response-distractor task (Galantucci, Fowler, & Goldstein, 2009; Kerzel & Bekkering, 2000), motivate the computational principles of both excitation and inhibition. DFT was chosen as the framework for building the models presented in this chapter and in Chapter 6 in large part because these principles are at the heart of DFT, as will be discussed in more detail in section 5.3.2.5. In addition, this particular framework has been chosen because DFT models of movement planning have been used to account for behavioral data across a wide variety of actions, e.g., saccades (Kopecz & Schöner, 1995), infant perseverative reaching (Thelen et al., 2001), and arm movements (Schöner, Kopecz, & Erlhagen, 1997). The model in Chapter 6 applies DFT to the movement planning of speech articulators. In that model, phonological representations embrace phonetic

detail, with phonetic parameters represented as activation fields. These fields evolve over time and determine the specific parameter settings of the utterance being planned. The model presented in this chapter is not one of movement planning, but rather of perception, following other researchers who have extended DFT to perception of, e.g., motion-pattern perception (Hock, Schöner, & Giese, 2003).

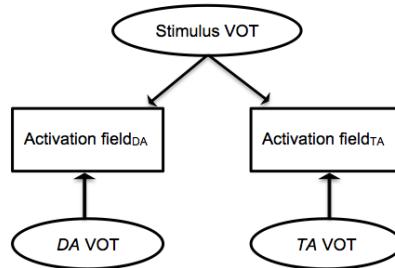
DFT has also been used by researchers to model other phonological processes, e.g., diachronic lenition (Gafos & Kirov, 2010; Kirov & Gafos, 2007), vowel-to-vowel coarticulation and dissimilation (Tilsen, 2007), and gestural drift in multilingual speakers (Tobin & Nam, 2010). The details of the DFT framework and the model of the phoneme-classification task are presented below.

### **5.3.2. Model of the phoneme classification task**

The main components of the model are shown in Figure 17. The model comprises two activation fields, one corresponding to each of the two potential responses in the task (here *ta* or *da*); and three sources of input, one corresponding to the stimulus in the experiment, and the other two corresponding to the speaker's representations of the potential responses on a given trial, i.e., VOT values for voiceless *ta* and VOT values for voiced *da*.

At the beginning of a trial, the TA and DA activation fields are set up in anticipation the two possible categories that the incoming stimulus may belong to, i.e., voiced or voiceless. When the stimulus is perceived, the relevant property of the

stimulus (here, VOT) serves as input to both of the activation fields. The stimulus input interacts with the activation levels of each field, causing the activation level of the field to which it is more similar to rise more quickly than the activation level of the



**Figure 17. Components of the model of the phoneme-classification task. Planning fields are shown in rectangles. Input to the fields is shown in the ovals.**

field to which it is dissimilar. This process continues until the more similar field stabilizes with an activation peak. The response on a given trial is determined to be the option represented by the field that has reached this stabilization point. The mechanisms by which these fields evolve and stabilize are explained in detail below before presenting the results of simulations of the phoneme-classification task.

### 5.3.2.1. Activation fields

Each activation field is defined by three axes, as illustrated in, e.g., Figure 18. One axis represents the possible values along the phonetic parameter of VOT. The second axis (here the vertical axis) represents the amount of activation associated with each VOT value. The VOT parameter field was defined on an arbitrary scale of -10 to 10, which were converted to the experimental VOT values by the linear transformation:  $(x + 3) \times 10$ , thereby representing a range of VOT values from -70 to

130 ms. The third axis represents time. The field evolves over time based on the influence of various inputs to the field, according to the dynamics inherent to DFT, which will be explained in detail in the next sub-section. With sufficient input, an activation field eventually stabilizes with a peak of activation centered at some VOT value. An example of a field with activation levels for VOT values typical of a *ta* response are shown in Figure 18A. The model decides which phoneme is perceived based on whether the DA or TA field stabilizes first.

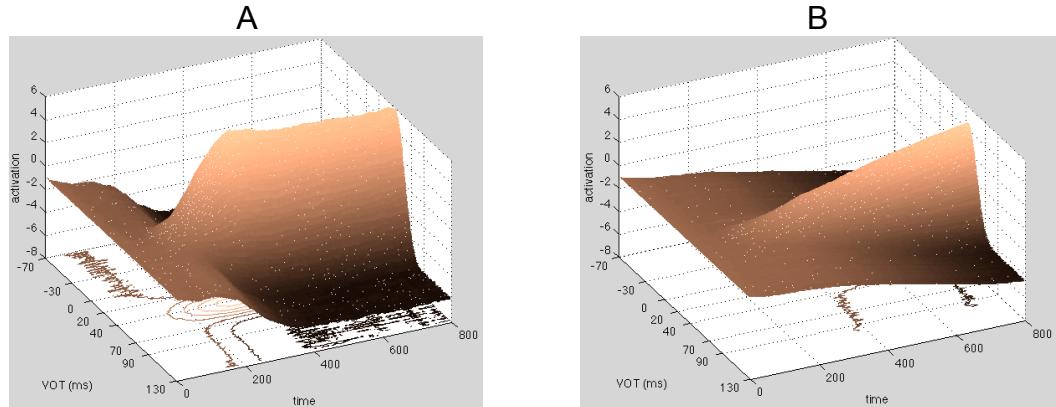
The use of VOT in milliseconds as the parameter defining voicing is not intended to assert that the voicing distinction is necessarily represented acoustically as VOT. The parameter of the representation could equally be cast as the relative phasing of an oral vocal-tract gesture with a glottal gesture (see, e.g., Browman & Goldstein, 1986; Browman & Goldstein, 1989). VOT was chosen to be as similar as possible to the experimental results being modeled. The claim here is also not that this phasing relationship realized as VOT is the sole acoustic cue in English to distinctions in voicing, which is not the case (see Cole, Kim, Choi, & Hasegawa-Johnson, 2007; Kingston & Diehl, 1994, and references therein). Further discussion of this point is taken up in section 7.3.4 of Chapter 7.

### 5.3.2.2. Model dynamics

The dynamical evolution of the activation fields is defined in (2).

$$(2) \quad \tau dA(x, t) = -A(x, t) + h + p(\text{input}_{\text{PRESHAPE}}(x, t)) + s(\text{input}_{\text{STIMULUS}}(x, t)) + \text{interaction}(x, t) + \text{noise}$$

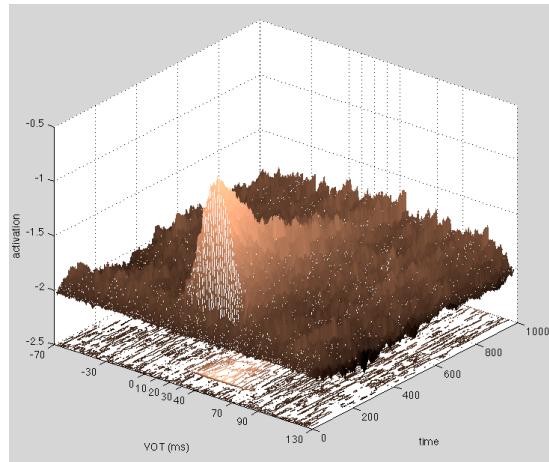
Each term in (2) is explained in detail below and in the sections that follow.  $dA(x, t)$  is the change in activation level  $A$  of parameter value  $x$  at time  $t$ . The rate of evolution of the field is controlled by  $\tau$ . Larger values of  $\tau$  result in slower evolution of the field, effectively “zooming in” on the dynamics to show how they evolve with a higher level of temporal discrimination. Figure 18 shows the evolution of two activation fields with an input value of 1 (i.e., 40 ms VOT). All aspects of the field evolutions are the same except for the value of  $\tau$  in each. Figure 18B with  $\tau = 240$  illustrates the evolution of the first temporal third of Figure 18A with  $\tau = 80$ . Figure 18 shows that it “takes longer” for the field to achieve a stable activation peak when  $\tau$  is larger. In all of the simulations in the model reported in this chapter,  $\tau$  is set to 160.



**Figure 18. Evolution of two fields with input at 40 ms VOT differing only in the value of  $\tau$ : (A)  $\tau = 80$  and (B)  $\tau = 240$ .**

$h$  is the resting level of the field, and is set to  $-3.25$  in the model. Uniform random noise is added to introduce stochastic behavior into the system (see section 5.3.2.6), adding a varying amount of activation at each time step. Therefore, without

any input,  $dA(x, t) = -A(x, t) + h$  means that the activation level for all values of  $x$  reverts to  $h + \text{noise}$  (around  $-2$ ). This is illustrated in Figure 19, where a short input with a maximum activation at 40 ms is introduced for a few time steps in the evolution. However, the input is of insufficient strength to engage the interaction term (the mechanism that introduces excitation and inhibition, see section 5.3.2.5) and stabilize with a raised peak of activation. The field therefore returns to the resting level.



**Figure 19. Activation field returning to resting level.**

The inputs defined in the next sub-sections are added to the field by the terms  $\text{input}_{\text{PRE-SHAPE}}(x, t)$  and  $\text{input}_{\text{STIMULUS}}(x, t)$ .  $p$  and  $s$  are weighting factors that indicate the relative strength of the pre-shape and stimulus inputs, respectively. A given input does not have to persist for the entire evolution of the field; its duration depends on the task. Therefore, at any time  $t$ , both, either, or neither input may be present.

### 5.3.2.3. Pre-shapes

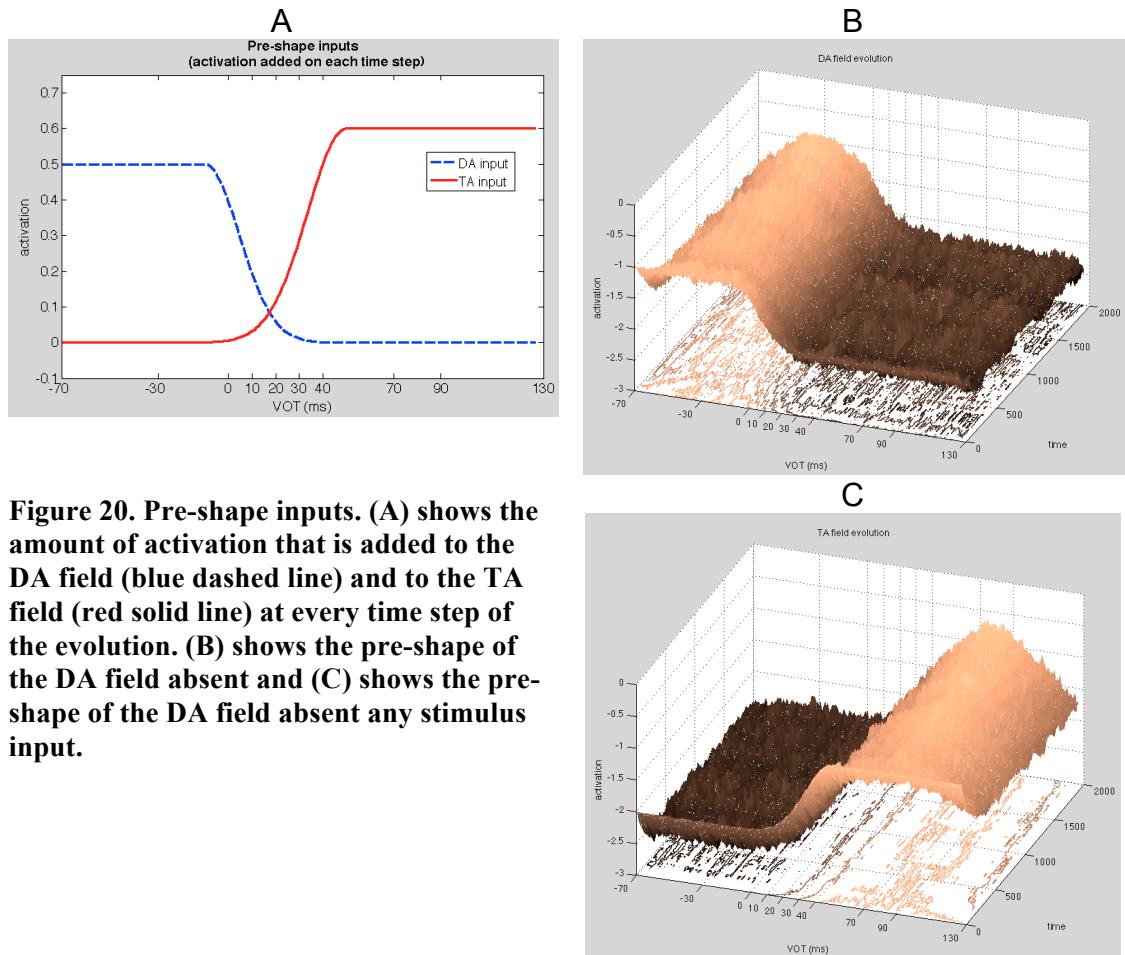
The first source of input to the activation fields is the pre-shape corresponding to each of the two possible responses on a given trial. The pre-shape of each field is based on the subject's representation of the parameter relevant for that field, here represented as VOT. The pre-shape input for one time step in the evolution of each field is shown in Figure 20A and derived from—but not defined by—the formula in (3).

$$(3) \quad \text{input}_{\text{STIMULUS}} = e^{-(x-\text{val})^2} / 2\sigma^2$$

The activations shown in Figure 20A obviously not normal distributions, as would be expected from (3). The distributions are modified from (3) as follows. *val* indicates the VOT value with the maximum level of the input activation distribution per the definition in (3), which is specified as  $-4$  for the DA field, corresponding to a VOT of  $-10$  ms, and  $2$  for the TA field, corresponding to a VOT of  $50$  ms. The height of the activation level is determined by the pre-shape weight *p* in (2), which is  $0.5$  for DA and  $0.6$  for TA.  $\sigma$  indicates the standard deviation of the distribution on which the pre-shape input is based, which is  $1.45$  for the DA input and  $1.65$  for the TA input. The reasons for the specific values chosen for the pre-shapes is addressed in detail in section 5.4.2.

For the DA input, the activation distribution for all VOT values to the right of *val* is as defined by (3), while the activation values for all VOT values to the left of *val*

were all the same as  $val$ . Similarly, the activation distribution for the TA field is left as defined in (3) from  $val$  and all VOT values leftward, and set to the activation level of  $val$  for all points to the right of  $val$ . The reason for structuring the distributions this way reflects both the phonemic boundary between syllable-initial voiced and voiceless stops in English and the nature of the categorization task. On either side of the phonemic boundary (here 20 ms VOT), stimuli are classified categorically



**Figure 20. Pre-shape inputs.** (A) shows the amount of activation that is added to the DA field (blue dashed line) and to the TA field (red solid line) at every time step of the evolution. (B) shows the pre-shape of the DA field absent and (C) shows the pre-shape of the DA field absent any stimulus input.

regardless of how far away from the phonemic boundary they are located, as Figure 16B shows for the classification results from Phillips et al. (2000). In effect, they represent the fact that any VOT value less than or equal to 10 ms is unambiguously *da*, any VOT value greater than 30 ms is unambiguously *ta*, and VOT values in between 10 and 30 ms are ambiguous. Alternative definitions for the pre-shapes are discussed in section 5.4.2.

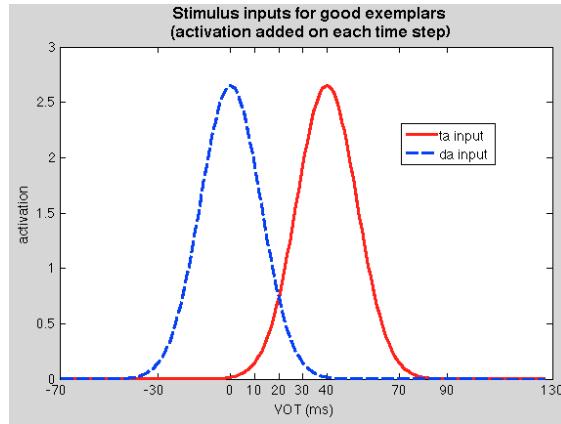
The pre-shape input is present at all time steps ( $t$ ) in the evolution of both fields. Figure 20B and C show the pre-shapes of the two activation fields. The pre-shape of each field is such that the activation level for the VOT values representing its category is notably higher than the VOT values representing the other category, but not so high that some VOT value stabilizes with an elevated peak of activity indicating a response.<sup>1</sup> The pre-shape persists at the levels shown in Figure 20 for as long as the fields evolved, regardless of any other input.

---

<sup>1</sup> Figure 20B and C show that the pre-shapes of the fields as they evolve over time differ slightly from the actual input distributions shown in Figure 20A. While the activation level of the input levels for DA and TA form a plateau on their respective high ends, these plateaus taper off toward the edge of the activation field as the field evolves. This is due to the specific implementation of the interaction term, which is defined in (4) and described in section 5.3.2.5. In short, one effect that the interaction term has on the field is that when some value  $x$  is above threshold, it raises the activation level of  $x$  values within some distance ( $\sigma_w$ ) by a set amount ( $w_{excite}$ ). Since every  $x$  value has this effect on its neighbors, the activation level of a given  $x$  value gets a boost of activation from each of its neighbor that is within distance  $\sigma_w$  and is over threshold. As an  $x$  value gets closer to the edge of the activation field than  $\sigma_w$ , it has fewer neighboring  $x$  values boosting its activation level. As a result, the activation level tapers off at the edge of the field. This effect of the specific implementation can affect the behavior of the model if inputs are near the field edges. Therefore, no such inputs are used in any aspect of the model other than the pre-shapes, where this property of the implementation has no critical impact on the behavior of the model.

#### 5.3.2.4. Stimulus input

The other source of input to the activation fields is the VOT of the auditory stimulus on the trial, the experimental manipulation of which was the crucial property for the phoneme classification. On a given trial in the model, the same input is introduced to both the DA and TA fields. The stimulus input on all trials starts at time step ( $t$ ) 100, and lasts 225 time steps. The inputs for a single time step in the evolution of the field are shown in Figure 21.



**Figure 21. Inputs of good-exemplar stimuli for *ta* with a VOT of 40 ms (solid red line) and *da* with a VOT of 0 ms (dashed blue line).**

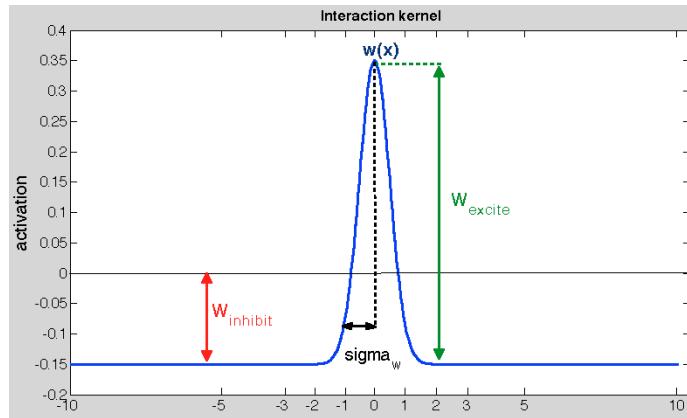
Unlike the pre-shapes, the stimulus inputs are normal distributions, defined by (3). The values for  $val$  are  $\{-3, -2, -1, 0, 1\}$ , corresponding to the experimental stimulus VOT values of  $\{0, 10, 20, 30, 40\}$  ms, respectively. The height of the activation level is determined by the stimulus weight  $s$  in (2), which is 2.65 for all inputs.  $\sigma$  indicates the standard deviation of the input distribution, which is 1.25 (i.e., 42.5 ms) for all inputs.

### 5.3.2.5. Interaction term

The interaction term,  $\text{interaction}(x, t)$ , is the “engine” that drives the evolution of the activation field through local excitation and lateral inhibition. How and to what degree the interaction term  $w(x)$  effects the evolution of the field is defined in (4).

$$(4) \quad w(x) = w_{\text{excite}} e^{-(x^2/2\sigma_w^2)} - w_{\text{inhibit}}$$

The inhibition term  $w(x)$  defines how much activation is added to each value of parameter  $x$ , at one time step  $t$  in the evolution of the field.  $w(x)$  is illustrated in Figure 22. For each value of  $x$  in the activation field, two things happen. First, the activation level of all values of  $x$  are reduced by some amount, determined by  $w_{\text{inhibit}}$ , which is 0.15 in the model presented here. This is the source of within-field lateral inhibition. Second, the maximum value of  $w(x)$ , which is 0.35, is added to the activation level of  $x_i$ , where the maximum of  $w(x)$  is determined by  $w_{\text{excite}}$  (which is 0.5) minus  $w_{\text{inhibit}}$ .

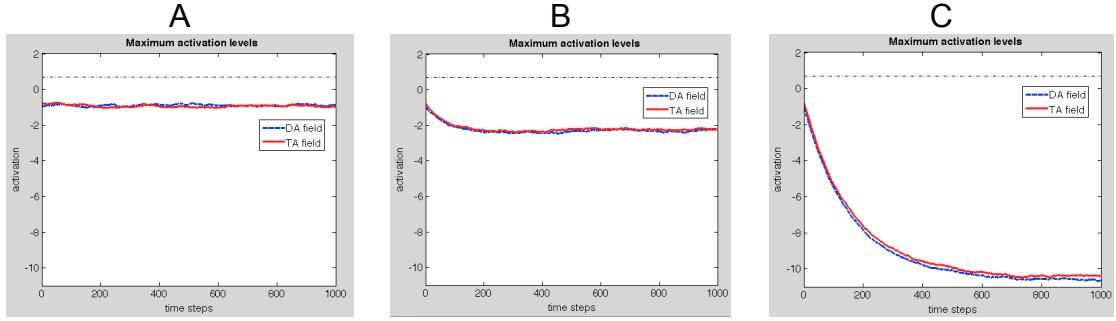


**Figure 22.** The interaction term  $w(x)$ , showing the values of (4) used in model.  $w_{\text{inhibit}} = 0.15$ ,  $w_{\text{excite}} = 0.5$ , and  $\sigma_w = 1.75$ . The units of the  $x$  axis are the arbitrary  $-10$  to  $10$  of VOT used in the model.

Activation is also added to values of  $x$  within a certain distance of  $x_i$ , where the amount of activation added to  $x_j$  decreases with distance from  $x_i$ , as determined by the width of  $w(x)$ . The width of  $w(x)$  is determined by  $\sigma_w$  (which was 1 in the model presented here). This is the source of local excitation. The excitation term can add activation to the field iff  $w_{excite} > w_{inhibit}$ . If  $w_{excite} \leq w_{inhibit}$  (and  $w_{inhibit} > 0$ ), then the interaction term will only introduce inhibition. The interaction term only induces changes in the field when the activation level  $u$  of some value of  $x$  approaches a soft threshold ( $\theta$ ).  $\theta$  is modeled as the sigmoid threshold function defined in (5).

$$(5) \quad f(u) = \frac{1}{1 + \exp[-\beta(u - \theta)]}$$

$\beta$  is the slope of the function and determines the extent to which activation levels that are below  $\theta$  influence the interaction (Erlhagen & Schöner, 2002, see caption of their Figure 4 and their Appendix A for more details). The higher the value of  $\beta$ , the lower the activation level of a given  $x$  location can be and still contribute to the interaction term. In other words, the maximum activation level of a field does not have to be at  $\theta$  in order for the interaction term introduce local excitation and lateral inhibition to the field. However, the further away the activation level of a location  $x$  is from  $\theta$ , the less it contributes to the interaction. This is illustrated in Figure 23.



**Figure 23. Maximum activation levels of the TA and DA fields with the same pre-shape input only (i.e., without any stimulus input) with three different values for  $\beta$ : 1.5 (A), 0.5 (B), and 0 (C).  $\theta$ , indicated by the dotted line, is the same (0.7) in all cases. The initial maximum activation level of both fields in all cases is set to about  $-1$  in order to illustrate the change in the field activation levels across cases.**

The value of  $\theta$  in all three conditions illustrated in Figure 23 is set at 0.7, i.e., the value of  $\theta$  in the model presented here, and is shown as a dashed line. None of the activation maxima in any of the three conditions ever reaches  $\theta$ . In Figure 23A,  $\beta$  is set at 1.5, which is also the value used in the present model. The maximum activation levels shown in Figure 23A therefore reflect the fields depicted in Figure 20B and C, and the maximum activation of both fields stays at roughly  $-1$ . Figure 23B shows that by changing the value of  $\beta$  from 1.5 to 0.5, the resting level of the fields drops to about  $-2$ . By reducing  $\beta$ , the activation values based on the pre-shape input alone contributes less to the field interaction compared to when  $\beta$  is greater. This reduces the amount of local excitation that the pre-shape inputs can generate, and the maximum level of the fields drops as a result. Figure 23C shows that the field maxima drops even further when  $\beta = 0$  because the interaction term does not have any effect on the field at all until the activation level of some  $x$  value actually reaches  $\theta$ . The sigmoid function (5)

thereby introduces non-linearity to the dynamics of the system, in that incremental changes in activation levels have a non-uniform effect on the system.

### 5.3.2.6. Noise

Noise is added to the evolution of the field to introduce stochastic behavior to the model. At each time step in the evolution, a random amount of activation is added to the activation level of all  $x$  values as defined in (6).

$$(6) \quad \text{noise} = v_x \times (\zeta \times \text{weight}_{\text{noise}})$$

$v$  and  $\zeta$  are each random numbers between 0 and 1.  $\text{weight}_{\text{noise}}$  is set to 7 in the model. Therefore, the amount of noise added to each  $x$  value ( $v_x$ ) varies within a given time step  $t$ , and the scaling factor of the noise varies from time step to time step. Since the means of  $v$  and  $\zeta$  are both 0.5, the minimum amount of noise added on a give time step is 0 activation units, the maximum is 7, and the mean is 1.75.

### 5.3.2.7. Summary

In summary, the model of the phoneme-classification task contains two activation fields that serve as detectors, one attuned to *ta* input and another to *da* input. In response to the stimulus input on a given trial, a peak of activation rises in one or both of the fields and ultimately stabilizes. A stabilized field indicates the categorical percept on that trial. The stimulus input interacts with the pre-shapes of each field depending on the compatibility of the stimulus VOT value and the pre-shape. Incompatible inputs to a field, that is, those that are far away from each other along the

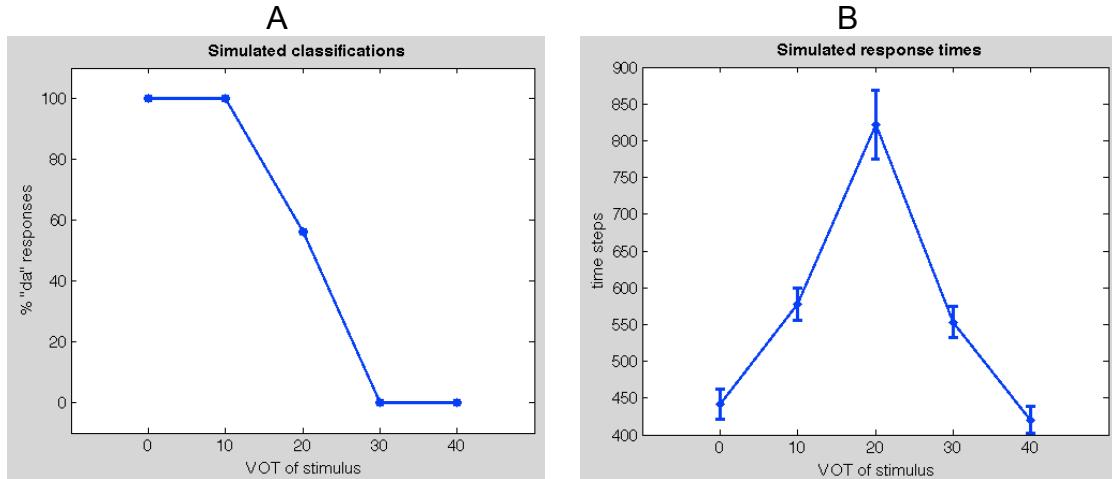
$x$  axis, inhibit each other as a result of the interaction term  $w(x)$ . Compatible inputs, that is, those that are near each other along the  $x$  axis, excite each other as a result of the same interaction term. These effects of local excitation and lateral inhibition affect the rate at which the activation peaks rise and achieve stabilization. The rate of activation rise is correlated with RTs in the experiment. The next section presents simulations of the various conditions in the experiment and illustrates how the interactions between the various stimulus inputs and the field pre-shapes account for the response times and classifications.

### 5.3.3. Experiment simulations

The phoneme classification task was simulated using the model described above. The solution to the equation in (2) was simulated with a MATLAB (R2011a, The MathWorks, Natick, Massachusetts) script using a step-wise Euler method (see, e.g., Higham, 2001). The MATLAB scripts that were used to simulate the single trials and the experiment can be found in Appendix A and Appendix B, respectively.

500 trials were simulated, with 100 trials for each of the five VOT values of the stimuli used by Blumstein et al. (2005): 0, 10, 20, 30, and 40 ms. The actual model parameter values used in the DFT model were  $\{-3, -2, -1, 0, 1\}$  (see section 5.3.2.1). On each trial, the TA and DA fields started with activation levels as described in section 5.3.2.3. The input corresponding to the VOT value of the stimulus being simulated on that trial ( $val$ ) was introduced at time step 100, and was the same for both

fields. Both fields evolved until one of the fields reached a criterion level ( $\kappa$ ). The criterion level was chosen to indicate that the activation level of some VOT value in one field had reached a sufficient level that that field would stabilize. Figure 25 shows that the fields stabilized around an activation level between 5 and 6 activation units, so a  $\kappa$  of 5 was used. Which field reached criterion first was logged as the classification response on that trial, as was the time step in the evolution when this field reached criterion. This time step minus 100 (the starting time step of the stimulus input) served as the simulated RT for that trial.

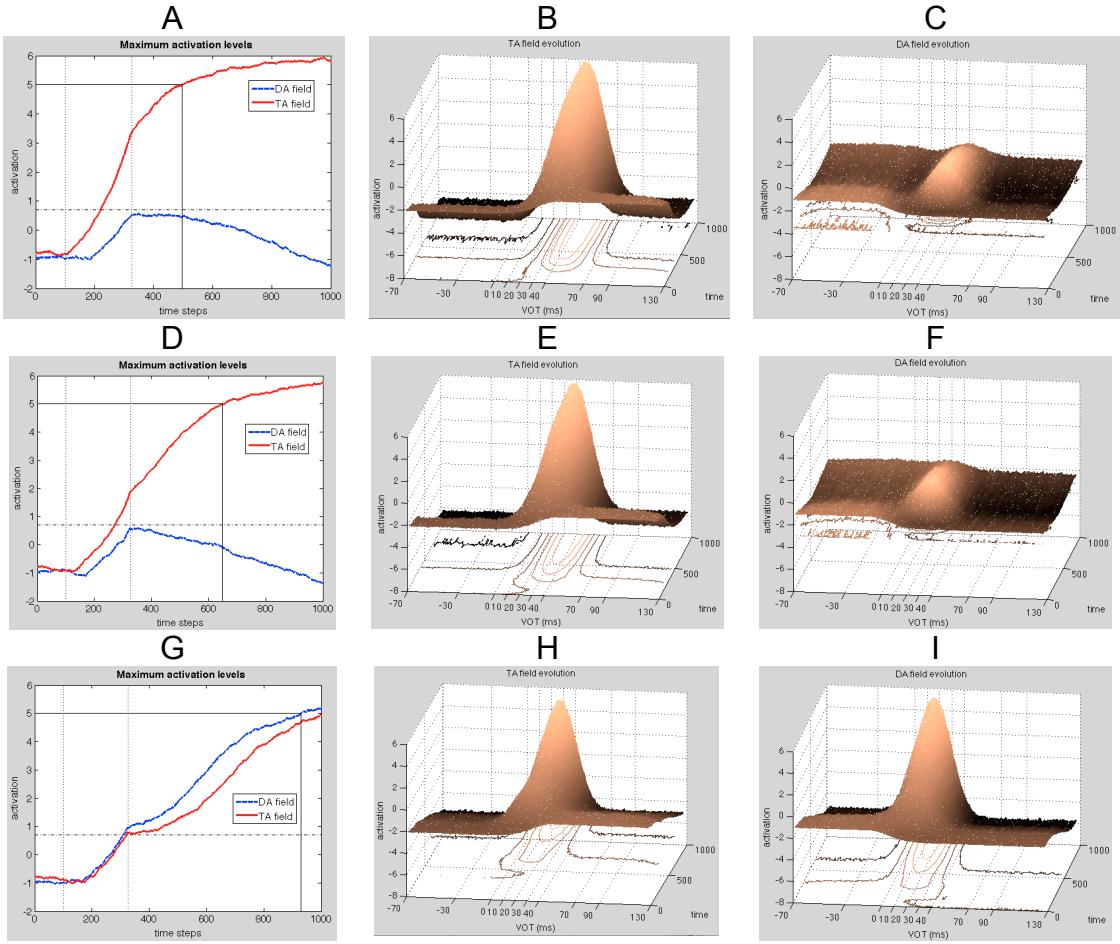


**Figure 24. The results of the simulations of 500 trials of the phoneme classification task, with 100 trials for each stimulus VOT value. (A) shows the classification results, and (B) shows the response times for the same trials. Compare to the actual experimental results from Blumstein et al. (2005) shown in Figure 16A.**

Figure 24 shows that both the classification and RT results from the simulations were qualitatively similar to the results obtained by Blumstein et al. (2005). Figure 24A shows the percentage of the trials where the response was classified as *da*. These classifications show the same qualitative pattern as the

experimental results, showing sensitivity to the phoneme VOT boundary. The 0- and 10-ms VOT stimuli were both categorically classified as *da*, and the 30- and 40-ms VOT stimuli were both categorically classified as *ta*. The classifications for the 20-ms VOT stimulus were ambiguous, with the stimulus being classified on about half of the trials as *da* and on the other half as *ta*. Figure 24B shows that the RT results from the simulations also qualitatively replicated the experimental results, with RT being modulated by category goodness. The good exemplars (0- and 40-ms VOT stimuli) had the fastest RTs, and the marginal exemplars (10- and 30-ms VOT) had slightly longer RTs. The ambiguous 20-ms stimulus had the longest RTs.

Figure 25 depicts the evolutions of the activation fields in representative trial simulations from three types of trials: one with a good exemplar stimulus (40 ms VOT, shown in the first row of graphs), one with a marginal exemplar stimulus (30 ms VOT, shown in the second row), and one with an ambiguous exemplar stimulus (20 ms VOT, shown in the second row). The next sub-sections explain how the dynamics of the trials shown in Figure 25 yield the results shown in Figure 24. The examples illustrated with Figure 25 all use various *ta* inputs to illustrate the dynamics of the model. The behavior of the model is qualitatively the same for *da* inputs as the design of the model is symmetrical, that is, a good exemplar stimulus of *da* affects the DA and TA fields in the same way that a good exemplar stimulus of *ta* affects the TA and DA fields, respectively.



**Figure 25.** Shows the evolution of the activation fields for simulations of three categories of stimuli in the phoneme-classification task. Each row of graphs shows the fields from one trial simulation of a stimulus from each exemplar category. The first row shows one simulated trial with a stimulus of the good exemplar category (40 ms VOT). The second row shows one simulated trial with a stimulus of the marginal exemplar category (30 ms VOT). The third row shows one simulated trial with a stimulus of the ambiguous exemplar category (20 ms VOT). The first column of graphs shows a comparison of the maximum activation level of the TA field (solid red line) and DA field (dashed blue line) in the evolution of each simulation. The vertical dotted black lines show the duration of the stimulus input, from time steps 100 to 325. The dotted black line shows the within-field interaction threshold ( $\theta = 0.7$ ). The horizontal solid black line shows the criterion activation level ( $\kappa = 4.75$ ) that was used to determine the response on the trial. The vertical solid black line shows the time step at which one or the other field reached  $\kappa$ , signifying the RT on that trial. The second column of graphs shows the evolution of the TA field in each simulation. The third column of graphs shows the evolution of the DA field in each simulation.

As noted in section 5.3.2.3 above, both activation fields start out with pre-shapes that reflect the subject's representation of the VOT corresponding to the two voicing categories that differentiate the two possible responses. Figure 25B and C show that one side of each field is higher than the other from time steps 1 to 100 in (these are the same as the pre-shape shown in Figure 27A and B, but shown here on a scale with a larger range of activation values on the vertical axis). The activation field matching the stimulus (in this example, the TA field) has a raised plateau of activation with a left edge of 40 ms, corresponding to a range of voiceless VOT values (Figure 25B, E, H). The pre-shape of the non-matching field (here DA) has a plateau of activation with a right edge of 0 ms, corresponding to a range of voiced VOT values (Figure 25C, F, I). The maximum value of the peak of activation for both pre-shapes is about -1, as can be seen in Figure 25A, D, G.

### 5.3.3.1. **Good-exemplar stimuli**

The simulation of a representative trial with a good exemplar stimulus is illustrated in Figure 25A, B, C. Figure 25B shows that when the stimulus input is introduced at time step 100, a peak of activation having a *val* equivalent to 40 ms (i.e., 1) builds at the left side of the pre-shape for the TA field. Figure 25A shows that from time step 100 through the duration of the stimulus input, the maximum of this peak rises due to a combination of augmenting inputs of the pre-shape and the stimulus with the local excitation introduced by the interaction term. The effect of the interaction

term alone can be seen clearly after time step 325: the maximum activation level of the TA field continues to rise after the stimulus input has ended. This illustrates the inherent stabilization in DFT. The values around the peak continue to locally excite each other to a greater degree than the lateral inhibition lowers the activation level of the field (recall from section 5.3.2.5 that the interaction term reduces the activation level of the entire field, not just the level of distant  $x$  values). Local excitation introduced by the interaction term moves the field to a stable attractor state, at which point the maximum peak of activation levels off.<sup>2</sup> The stable state is achieved when the effect of activation increase from excitation offsets equally the effect of lateral inhibition. The field will stay at this level indefinitely until some inhibitory input is introduced. In the simulations, the criterion value ( $\kappa$ ) of 5 was used to indicate that a field that crosses it will inevitably reach that stable state and be the percept on that trial. In this simulated trial, the TA field reaches  $\kappa$  at time step  $\sim 350$ , so a RT of  $\sim 250$  is recorded.

The effects of inhibition in the dynamics of the model are illustrated by the behavior of the non-matching activation field (here the DA field). Figure 25C shows the pre-shape plateau along the voiced VOT values before the stimulus input. With the introduction of the stimulus input (again, with a peak of about 40 ms, the same as for the TA field), the stimulus-introduced values start to rise above the resting level, but

---

<sup>2</sup> The reason that the slope of the activation rise of the TA field slows down after time step 325 is that the rise between time steps 100 and 325 reflects both stimulus input and local excitation, while after time step 325 it reflects only local excitation.

do so at a longer rate than in the TA field because there is no augmenting input from the pre-shape and because of the inhibition introduced by the incompatible pre-shape. This can be seen by comparing the slopes of the red and blue lines in Figure 25A.

Another effect the interaction term can be seen in Figure 25C. As the stimulus input is introduced, a decrease in the activation level of the pre-shape in the voiced range of the  $x$  axis can be seen. This is due to the inhibition of the field introduced by the stimulus input, which is incompatible with the voiced pre-shape values. This inhibition is introduced even though the maximum activation level of the DA field never reaches  $\theta$ , as can be seen in Figure 25A. This demonstrates the “soft” nature of the sigmoid threshold function.

When the  $ta$  stimulus input to the DA field stops, the activation level immediately starts to drop back to the resting level of the field because the maximum activation level of the field has not gotten close enough the interaction threshold ( $\theta$ ) to maximize the effect of the interaction term and introduce sufficient local excitation. In trials with good-exemplar stimuli, the non-matching field has no chance of reaching the selection criterion ( $\kappa$ ). Therefore, in all simulations whose results are shown in Figure 24A, the field that is chosen as the response on a given trial is always the one corresponding to the category of the stimulus.

Another effect of lateral inhibition (although not one that has a material influence on modeling the experimental results) can also be seen in Figure 25B. At the beginning of the trial, the pre-shape of the TA field has an activation plateau starting

at around 30 ms and continuing to the rightmost edge of the field, as explained in section 5.3.2.3 (and see also footnote 1). As the activation peak resulting from the stimulus input rises, the activation at those  $x$  values starts to inhibit the activation level of all other  $x$  values in the field, including those that are also in the voiceless range as defined by the pre-shape. This is evident by the activation level for VOT values greater than about 70 ms dropping off as the peak of activation around 40 ms rises.

In summary, for the good exemplar condition, the combination of the maximally compatible pre-shape and stimulus inputs and the local excitation introduced by the interaction term resulted in the fastest rise in activation level in the matching field compared to the other conditions (marginal and ambiguous), where inhibition slowed down the rate of rise of the winning field (the causes of which are discussed below). This rapid rate of rise resulted in the model simulating the fastest RTs for the good exemplar condition, shown in Figure 24B.

### 5.3.3.2. Marginal-exemplar stimuli

A representative trial of the marginal exemplar condition is shown in Figure 25D, E, F. In this trial the  $val$  of the input stimulus is 30 ms, which is closer to the phoneme boundary of 20 ms. The TA field reaches  $\kappa$  at about time step 650 (i.e., a simulated RT of about 550) on this trial, which is notably later than the RT in the good exemplar condition. The differences between the field evolutions in this condition and in the good exemplar condition are slight, but they have an effect on the simulated

RTs. The most material difference is that in this condition the peak of the stimulus input is further away from the plateau of the pre-shape of the compatible field, which for the TA field starts at a VOT value of about 40 ms (see Figure 20). As a result, the slope of the line indicating the rate of rise of the activation maximum in the TA field is less in this condition than in the good exemplar condition. This can be seen in Figure 25D: the TA field maximum activation is at only about 2 activation units when the stimulus input stops, as opposed to the good exemplar condition where it rises to about 3.5 by that time (compare Figure 25A). This slower rate of rise is due to the pre-shape not augmenting the stimulus input as much as it does in the good exemplar condition, plus slight inhibition of the input activation peak from the  $x$  values of the pre-shape activation. With the leftward shift of the input peak away from the pre-shape plateau (compared to the peak in the good-exemplar condition), there are  $x$  values in the pre-shape that are sufficiently far away from the input peak that they slightly inhibit the input peak without adding any local excitation to it. The input peak still reaches a sufficiently high level of activation that it continues to rise due to the local excitation introduced by the interaction term, but it achieves that stable state later than in the good exemplar condition. Thus, the simulated RTs are longer in this condition than in the good exemplar condition.

The evolution of the mismatching field (Figure 25F) is qualitatively the same in this condition as in the good exemplar condition. This can be seen by comparing the dashed blue lines in Figure 25D and A. In this condition again, the stimulus input does

not reach a sufficient activation level to allow it to self-stabilize, and it drops back down to resting level after the stimulus input has stopped. As in the good-exemplar condition, the mismatching field never reaches  $\kappa$ , so it is always the matching field that stabilizes and dictates the categorical classification of these stimuli.

### 5.3.3.3. Ambiguous-exemplar stimuli

Figure 25G, H, I show that the evolutions of a representative trial from the ambiguous exemplar condition are qualitatively different from the evolutions in the two conditions described above. Most notably, in this condition both the matching and mismatching fields reach  $\kappa$ , so the trial classification and RT are determined by which field reaches  $\kappa$  first, rather than by which field reaches it at all. In this simulated trial, the DA field reaches  $\kappa$  slightly before the TA field and thereby determines that the response was classified as *da* with an RT of  $\sim 825$  (i.e., the field reached  $\kappa$  at time step  $\sim 925$ ). In other simulated trials of the ambiguous exemplar condition, it is sometimes the TA field that reaches  $\kappa$  first. The noise term causes the classifications to be at chance between *da* and *ta* in the ambiguous case. The fields evolve roughly equally during the time when the stimulus input is present—the only differences between the evolution of the fields are due to those introduced by noise. Whichever field winds up with a slightly higher level of activation when the stimulus input stops will most likely reach  $\kappa$  first, and thus determine the classification and the RT.

The reason why RTs are slowest in this condition should be clear from the discussion of the marginal condition above. The location of the stimulus input peak to the TA field (with  $val = 20$  ms) is further to the left of the TA field pre-shape plateau than the input in the ambiguous condition. The input peak therefore gets even less augmenting activation from the pre-shape, and it is inhibited even more by the pre-shape of the field since there are even more  $x$  values to the right of input peak that inhibit but do not excite the input peak. As a result, the rate of rise of the activation maximum of the field is slower than in the marginal condition, with the field maximum activation barely reaching 1 when the stimulus input stops. The level of activation of the peak is sufficient to reach a stable state, but it takes longer for the interaction term to raise the peak to that level since it starts out lower than in the other conditions. The situation is the same in the DA field evolution, since the input in this ambiguous case is equally far from the edge of the pre-shape plateau of the DA and TA fields.

#### 5.4. Discussion

This section discusses various aspects of the model and its simulations of the experimental results. Sub-section 5.4.1 discusses the results of the model. 5.4.2 motivates the pre-shapes used in the model. 5.4.3 highlights the key aspects of DFT as illustrated in the model.

### **5.4.1. Results of the model**

The main result of the simulations presented here is that the model simulations replicated the qualitative pattern of both the classification and RT results reported in the literature for this task. It was not the goal of the model to replicate the fine quantitative differences between conditions. There are two main reasons for focusing on the qualitative matching. The first is that the quantitative differences in the experimental results were not always statistically significant, e.g., the difference in classifications of 0- and 10-ms stimuli in the Blumstein et al. (2005) study (Figure 16A), and therefore should not be given much weight. The second is that slight differences in results from one study were not always replicated in another. For example, the classification results of the present model simulation were 100% categorical for non-ambiguous stimuli (see Figure 24A), while the classification results from Blumstein et al. (2005) showed a bias toward *ta* responses: their 0- and 10-ms stimuli were sometimes classified as *ta*, but their 30- and 40-ms stimuli were never classified as *da* (Figure 16A). However, in the experiment run by Phillips et al. (2000), 0- and 10-ms stimuli were never classified as *ta*. Similarly, whether a stimulus was truly 50-50% ambiguous seems to depend to a degree on what stimuli are used in the experiment. In the Blumstein et al. (2005) study, the 20-ms stimulus is the ambiguous one, while in the Phillips et al. (2000) the 20-ms stimulus was classified as *da* about 80% of the time and the 25-ms stimulus was 50-50% ambiguous. The source of these differences could have been a result of several factors: the number of stimuli

in the experiment, the distance between the VOT values within an experiment, experimental setup, and/or differences between subject populations.

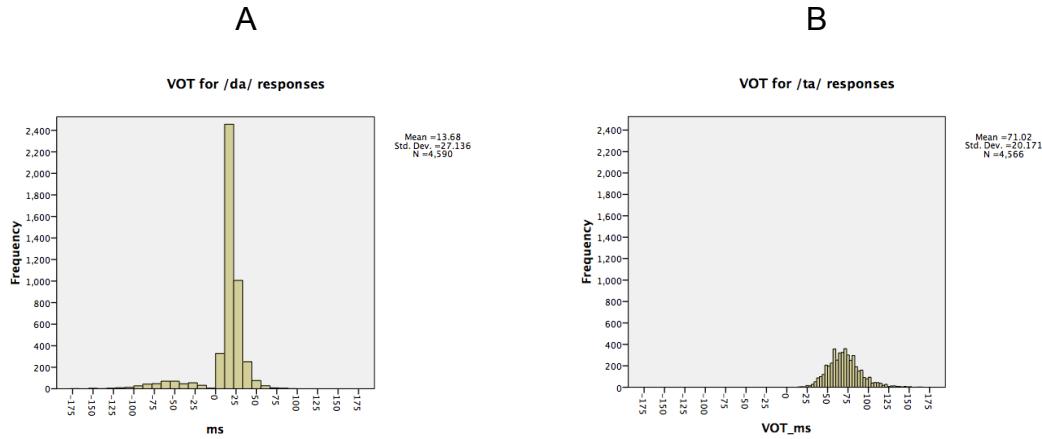
#### 5.4.2. Pre-shapes

In the model a field stabilizing at an elevated level of between 5 and 6 activation units indicates a response choice, which does not happen before a stimulus is presented. The weight value of the pre-shape input ( $p$ ) in (3) on each time step ( $t$ ) was chosen to be as large as possible such that the fields would maintain a peak of activation reflecting their VOT voicing category but would not rise sufficiently high as to engage the interaction term to the point that the field would stabilize with a peak of activation indicating a specific VOT value.  $r$  was set to 0.75 for both fields.

It is not a requirement of the DFT framework that inputs to fields be normal distributions like the input defined in (3), and indeed the pre-shapes were not modeled as normal distributions. The pre-shape input in the model of this task is intended to reflect a speaker's representation of the phonetic parameter of the field (VOT) for the category that the field represents, for the purpose of classifying a stimulus along that continuum. This sub-section motivates the implementation of the pre-shapes as discussed in section 5.3.2.3.

The pre-shape input should be based on the linguistic experience of the speaker. It would seem reasonable then to have the VOT pre-shape distributions reflect actual productions. Figure 26 shows the VOT values for all of the productions

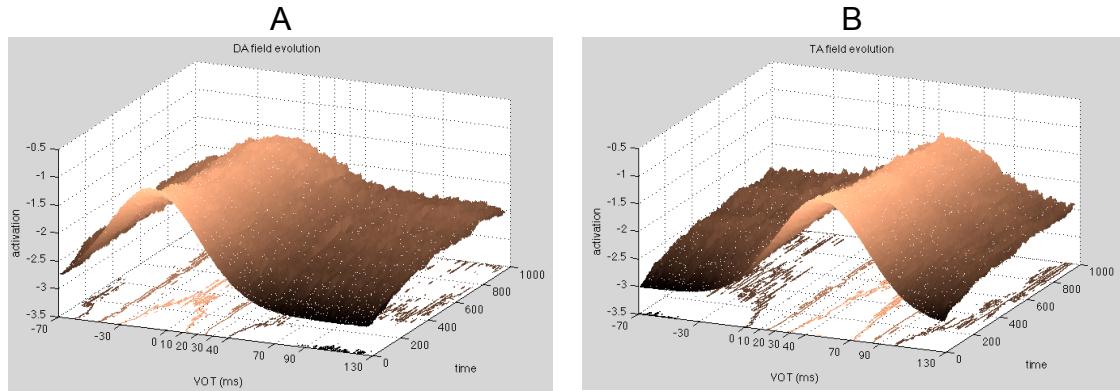
of *ta* and *da* from the experiments reported in Chapter 3. Figure 26A shows that the VOT distribution for voiced *da* is bimodal, with one mode representing the vast majority of utterances with a VOT between 12 and 25 ms, and a much wider distribution with far fewer tokens showing pre-voicing (i.e., negative VOT). Figure 26B shows that VOT for voiceless *ta* productions were uni-modal and very close to normally distributed, with a mean VOT of 71 ms.



**Figure 26. Histograms of VOTs of *da* (A) and *ta* (B) productions of all speakers from the experiments reported in Chapter 3.**

Assuming for the moment that these utterances are representative of American English VOT productions in general for these consonants, it is possible to make pre-shape inputs that reflect the actual production distributions more closely. Figure 27 shows the VOT pre-shape inputs as normal distributions, as opposed to the pre-shapes used in the model as implemented in section 5.3, which will be referred to here as the “plateau” model (see Figure 20) to differentiate it from a model using normally distributed pre-shape input. This is a simplification of the actual VOT distributions

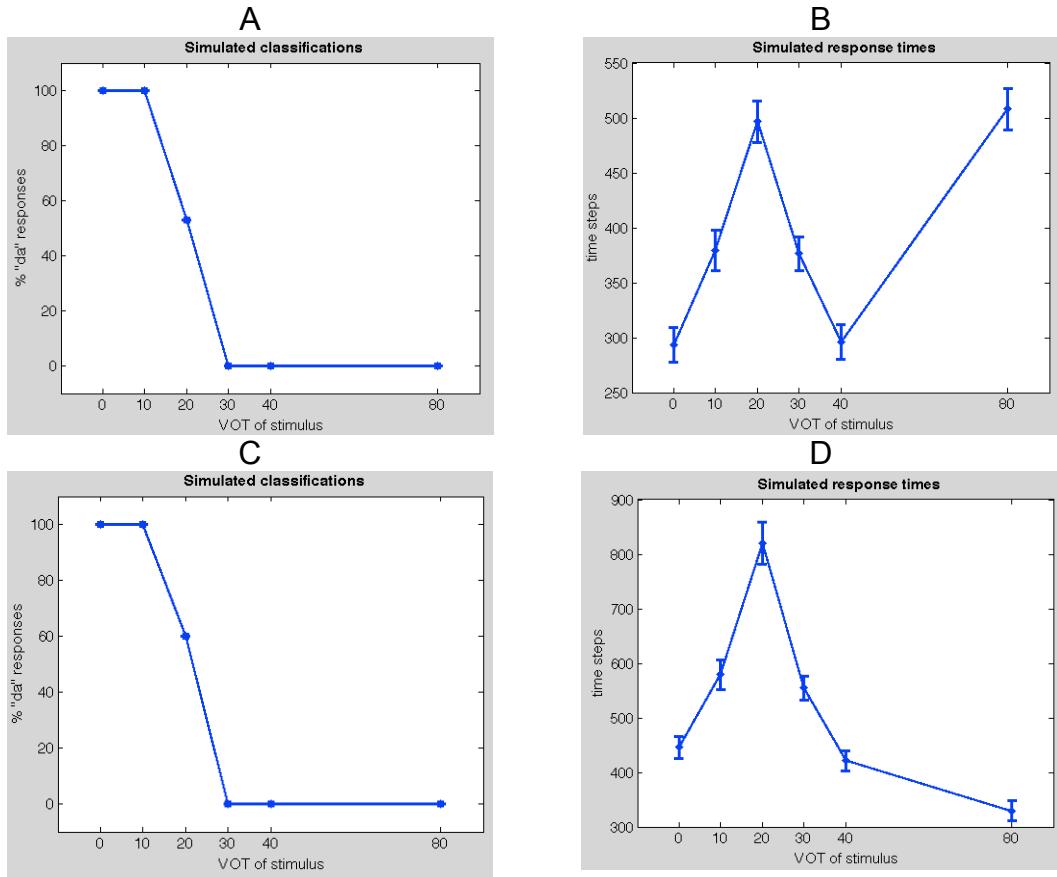
from the experiment (e.g., the distribution of  $da$  VOTs was bimodal in productions from the experiments), but even still it will serve to illustrate why this approach is undesirable. In this implementation, the TA pre-shape was defined by (3) with  $val$  being the same as in the model above (3, i.e., 50 ms), but  $\sigma$  was set to 3 (instead of 1) to allow for more  $x$  values to fall within range of the TA pre-shape. The DA pre-shape had  $val = -4$  (-10 ms) and  $\sigma = 3$ .



**Figure 27. Pre-shapes modeled as normal distributions as defined in (3) for the DA field (A) with  $val = -4$  and  $\sigma = 3$ . B) shows the pre-shape of the TA field with  $val = 2$  and  $\sigma = 3$ .**

The model simulations were run again using the pre-shapes shown in Figure 27, but this time a sixth stimulus with a VOT of 80 ms was included. As above, 100 trials were simulated for each stimulus VOT. All of the model variable values were the same as in the simulation reported above, with the exception of the stimulus strength  $s$ , which was set to 3 instead of 2.65. This change was needed to ensure that the input strength was sufficiently strong for the 80-ms stimulus so that one field or the other would reach criterion. The results of the simulation are shown in Figure 28A and B. Simulations using the model above including the 80-ms stimulus are shown in Figure

28C and D for comparison. Figure 28 shows that the model using the normally distributed pre-shapes did an equally good job of accounting for both the classification and RT data for the stimuli used in Blumstein et al. (2005) as the plateau model. Both models also classified the 80-ms stimulus categorically as *ta*, which is the result reported by Phillips et al. (2000), shown in Figure 16B.



**Figure 28. Model simulations including a *ta* stimulus with VOT = 90 ms. The top row shows the simulated classification (A) and RT (B) results for the model using normally distributed pre-shapes. The bottom row shows the simulated classification (C) and RT (D) results for the “plateau” model, i.e., having pre-shapes as defined in section 5.3.2.3.**

However, the model with normally distributed pre-shapes yielded problematic results for RTs to the 80-ms stimulus, which were as long or longer than for the ambiguous stimulus (Figure 28B). This is very different from the RTs reported by Phillips et al. (2000) for 80-ms stimuli, which were no different from the RTs for 40-ms stimuli in their experiment (Figure 16C). The reason for the long RTs for the 80-ms stimulus in the simulations with the normally distributed pre-shapes was that the stimulus input peak was as far away from the peak of the pre-shape (which is at 50 ms) as the ambiguous stimulus was. Therefore, the pre-shape peak actually inhibited the input peak introduced by the stimulus, causing the field to take longer to stabilize. Nevertheless, the model always classified the 80-ms stimulus as *ta* because the TA field pre-shape level of activation at 80 ms was higher than resting level and than the DA field at 80 ms, so the model was biased to having the TA field reach stabilization for longer VOT values.

Figure 28D shows that the simulated RTs for the plateau model, which has the RTs for the 80-ms stimulus being even shorter than the good exemplar stimuli. These RT results are also slightly different from results reported by Phillips et al. (2000), who found that the RTs to the 80-ms stimulus were no different from the good exemplar RTs. However, the RT data from Phillips et al. (2000) suggest that the RTs for the good stimuli are at a performance floor, i.e., subjects simply cannot reply faster than about 550 ms on average in this task. There is no mechanism in the present model to impose such a task performance floor, so the shorter RTs predicted by the model

based on stimulus properties are most likely moot. However, the problematic RT result provides a strong argument against using the normally distributed pre-shapes, since any such floor is not relevant to that case. The model in section 5.3 was therefore implemented as having pre-shape plateaus.

Another aspect of the pre-shapes that deserves some discussion is the use of slightly different weight  $p$  and  $\sigma$  values for the TA and DA pre-shapes (see section 5.3.2.3 and Figure 20A), which were both slightly higher for the TA pre-shape than for the DA pre-shape. Having those two parameters differ between the two pre-shapes resulted in the simulated RTs for the good *ta* stimuli being slightly shorter than for the good *da* stimuli (see Figure 24B). This reflects the numerical difference in RTs for the same stimuli reported by Blumstein et al. (2005) shown in Figure 16A. However, that difference was not significant, and was not replicated by Phillips et al. (2000). The differences are included in the present model simply to illustrate how differences in those variable values can result in differences in the model simulations.

#### **5.4.3. Comments on the dynamics of the model**

Since a primary goal of this chapter is to introduce DFT, this sub-section highlights a few important aspects of the framework as highlighted by the model. An important point to make about the settings of the model is that specific values of the various variables are not particularly meaningful on their own, and indeed there are other sets of variable values that would also yield qualitatively similar results to those

used here. It was instead the relative values taken in the context of the other variable settings that mattered most for the behavior of the evolution of the activation fields.

An inherent property of DFT is that activation fields have only two stable states: the resting level or one peak of activation that is induced by input and maintained by virtue of the interaction term. This stabilization is crucial to the model of the perceptual task presented here. Looking at the DA field evolutions in Figure 25A, D and G, it is clear that only when a field activation exceeds threshold  $\theta$  does it stabilize above the resting state of the field. The model here relies on this stabilization as the mechanism by which a percept is arrived at: no separate or higher-order decision-making mechanism is required. The criterion value  $\kappa$  is a computational convenience for indicating that a percept has formed by virtue of a field stabilizing.

There are three other aspects of the DFT formalism that are crucial in accounting for the model results: excitation, inhibition, and stochasticity. The computational principles of excitation and inhibition are at the core of the DFT interaction term. The combined effects of local excitation and lateral inhibition are the source of the RT differences in the model across the various stimulus conditions, as outlined in section 5.3.3. The introduction of noise in the evolution of the field (see section 5.3.2.6) introduces stochastic behavior that accounts for two aspects of the model simulations. First, this stochastic behavior is how the model produces a range of RTs for a given condition, rather than always yielding the same RT for a given stimulus. Second, the noise term accounts for the ambiguous classification of the 20-

ms stimulus, as described in section 5.3.3.3. The noise term is the source of the variation in which field reaches the criterion  $\kappa$  on a given trial, unlike the other conditions where only one field does so. Without the noise term, there would be no variation in the rate of evolution of the fields. If the pre-shapes of the two fields were completely symmetrical and the ambiguous stimulus was exactly half way between the two pre-shape edges, the two fields would always reach  $\kappa$  at the same time. In this case some other mechanism would have to be introduced in the model to choose a “winning” field. On the other hand, if some imbalance were introduced either in the pre-shapes (as was the case in the model as implemented) or in the location of the stimulus relative to the plateau edges, then the model would always classify the stimulus based on whatever bias was introduced by the asymmetry. The ambiguous condition in this experiment therefore illustrates the important role of the stochastic nature of the dynamics. The crucial role of the noise term in this model is consistent with other DFT models. For example, Hock et al. (2003) note that their model of visual motion pattern formation would not work without noise.

## 5.5. Summary and conclusions

The main purpose of this chapter was to introduce the computational framework of Dynamic Field Theory, which is used in the next chapter to model the experimental results presented in Chapter 3. DFT provides a set of computational mechanisms that define the evolution of activation fields based on excitation and

inhibition induced by various inputs. The dynamics of the field are driven by an interaction term that moves fields toward one of two stable states, either a rest state or a sustained peak of activation corresponding to some phonetic parameter value. This framework was used to model two results from a common phoneme-classification task, i.e., that classifications show categorical sensitivity to phoneme boundaries along a phonetic continuum while response times show gradient sensitivity to stimulus category goodness. The results of simulations by the model presented in this chapter are consistent with other theories of speech perception (e.g., TRACE of McClelland & Elman, 1986). The present model takes advantage of the inherent properties of DFT to simulate the results. No separate decision-making mechanism is required. Local excitation and lateral inhibition give rise to the response time differences, while field stabilization and stochasticity give rise to the classification results.

## **CHAPTER 6: A DYNAMICAL MODEL OF THE TIMECOURSE OF PHONOLOGICAL PLANNING**

### **6.1. Introduction**

This chapter presents a model of the timecourse of phonological planning, based on the response-distractor task used in the experiments reported in Chapter 3. The model is proposed as the normal mechanism by which the phonological parameters are set in speech production, with certain adjustments to accommodate the response-distractor task specifically. The model formalizes the link between speech perception and speech production as the phonological parameter values of a perceived utterance serving obligatorily as input to the normal process of the planning of phonological parameters for an utterance to be produced.

The model will be shown to account for the response time (RT) differences in the Congruent and Incongruent conditions of the two experiments reported in Chapter 3 (henceforth, “experiment 1” and “experiment 2”), where RTs were longer when a response and a distractor mismatched in articulator and voicing (e.g., *da-pa*, the Incongruent condition) than when they mismatched on only one of those parameters (e.g., *da-ba* or *ta-da*, the Congruent conditions). The results from the Congruent conditions in experiments 1 and 2 were qualitatively the same: RTs were both longer than in the Tone condition and shorter than in the Incongruent condition.

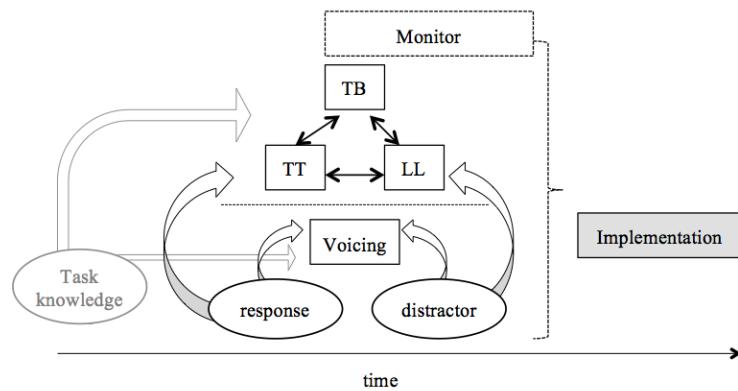
The model also yields two other results. First, the model accounts for findings from other studies using the same experimental task (Galantucci, Fowler, & Goldstein, 2009; Kerzel & Bekkering, 2000) that found RTs were fastest when the response and distractor were identical (in terms of phonological parameters) than when they mismatched on one parameter. The results for this identity condition were qualitatively different from other conditions in that only RTs in the identity condition were shorter than a neutral (tone) distractor. However, the identity condition need not be treated as a special case in this model: the result follows from the computational principles of inhibition and excitation, which are inherent to the computational framework as explained in Chapter 5 and motivated for these data in Chapter 4.

Second, the model accounts for an unexpected result from the experiments in Chapter 3. On any given trial in experiment 1, subjects could be 100% certain of the articulator of their response, but voicing was unpredictable. On the other hand, on any given trial in experiment 2 the articulator was unpredictable, while subjects could be 100% certain of the voicing of their response. While the general pattern of results with regard to (In)congruity were qualitatively the same across the two experiments, RTs were markedly longer (on the order of 50 ms) in experiment 1 than in experiment 2. That is, subjects responded more quickly if they knew the voicing of their response compared to when they did not. The model provides a principled account of this difference in RTs across experiments, and also makes novel predictions.

Section 6.2 presents the design of the model, enumerating its components. Section 6.3 defines the equations that control the dynamics of the model, and illustrates the behavior of the model in the various experimental conditions of the response-distractor task. Section 6.4 shows the results of simulations of experiments 1 and 2 from Chapter 3. Section 6.5 explains the variables that were used in the model, and their values. These properties of the model and results of the simulations are discussed in section 6.6 along with additional predictions of the model. Section 6.7 concludes.

## 6.2. Model components

The model is implemented using the computational framework of Dynamic Field Theory ("DFT", Erlhagen & Schöner, 2002; Kopecz & Schöner, 1995), the workings of which are explained in detail in Chapter 5. The functional components of the model are shown in Figure 29.



**Figure 29. Components of the dynamical computational model of phonological planning in the response-distractor task. Planning fields are denoted in rectangles with solid lines.**

**Inputs to the planning fields are shown in ovals. There is one input based on the phonological parameters of the required response, and another input based on the phonological parameters of a perceived auditory distractor. Given the nature of the task, the subject could anticipate certain aspects of the response on a given trial. These anticipated values also serve as input to the planning fields, labeled as “Task knowledge”. A Monitor function checks for when the activation level of the Voicing field and one articulator field (LL = Lower Lip, TT = Tongue Tip, TB = Tongue Body) reach a threshold value, at which point the values of the Voicing and winning articulator fields were selected to be sent to Implementation.**

The model includes four dynamical planning fields, inputs to these planning fields that determine the actual parameter values to be produced, a Monitor function that decides when all of the required values have been determined, and an Implementation component that executes the motor plans for the intended utterance based on the production parameter values determined by the model. Each of these components is described in more detail below.

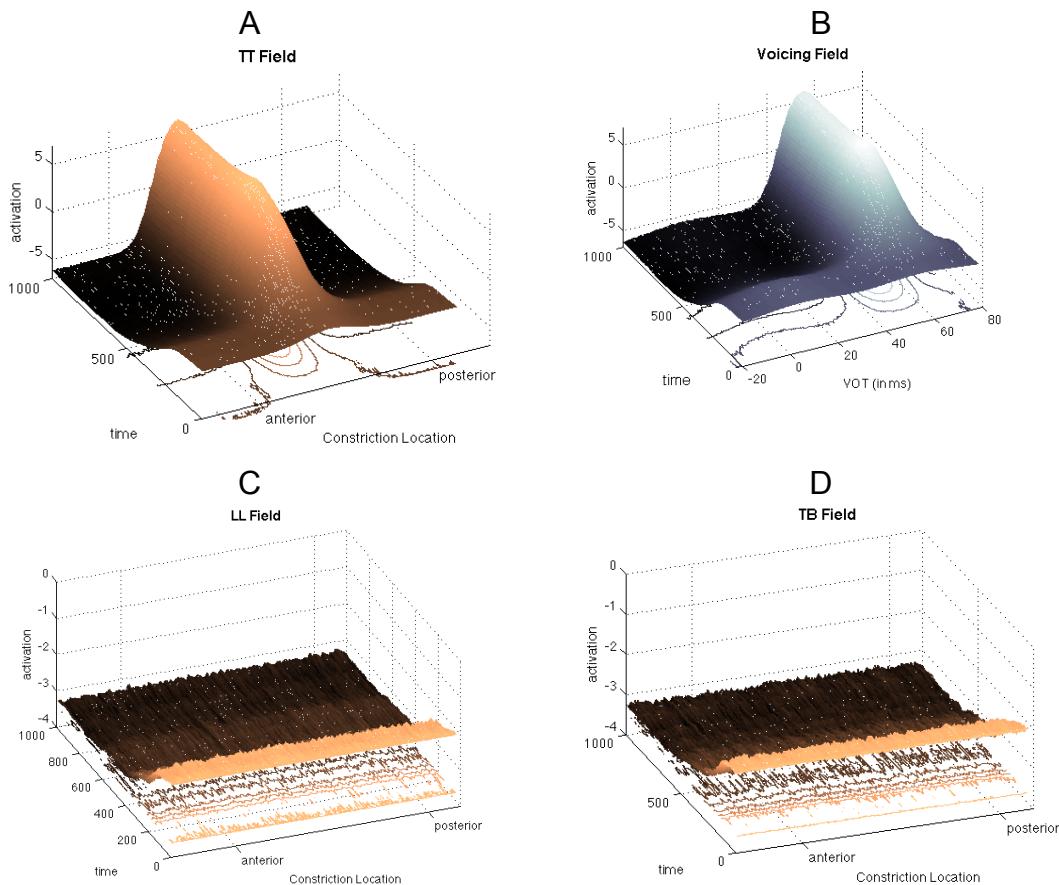
### **6.2.1. Planning fields**

There is a planning field for each of the phonological parameters that needs to be set for the utterance. The model is limited to fields for the parameters that were manipulated in the experiments in Chapter 3. There is one planning field for voicing, and one for each of the three primary oral articulators used in producing English stop consonants (Ladefoged, 1999), the lower lip (LL), tongue tip (TT), and tongue body (TB).

Each of the articulator planning fields is defined as detailed in section 5.3.2.1 of Chapter 5, with one axis representing the possible parameter values, one axis

representing the activation level associated with each possible parameter value, and a third axis representing time. The parameter axis in each field defines the constriction location for that articulator, represented as a continuum of more anterior to more posterior locations in the vocal tract relative to the possible constriction locations for that articulator. For the TB field, the constriction location values range from palatal (most anterior) to uvular (most posterior). For the TT field, the constriction locations range from dental (most anterior) to post-alveolar (most posterior). For the LL field, the constriction locations range from protruded (most anterior) to dental (most posterior). Since constriction location does not vary in the examples used in the model of this task, the input values for constriction location did not materially change in the simulations. The planning field for Voicing has a parameter axis represented as Voice Onset Time (VOT—see discussion in section 5.3.2.1 of Chapter 5 for discussion of the use of this measure to index the relative timing between the oral and glottal gestures of syllable-initial English stops).

Figure 29 shows the planning fields for a production of *ta*. The TT field shows a build-up of activation at the required constriction location (e.g., alveolar), and the Voicing field shows a build-up of activation at a VOT value of about 55 ms, corresponding to the voiceless, syllable-initial stop.



**Figure 30. Planning fields for *ta*.** The Tongue Tip (TT) field (A) stabilizes with a peak of activation at a Constriction Location value corresponding to alveolar. The Voicing field (B) stabilized with a peak of activation at a value corresponding to voiceless (~ 55 ms). The Lower Lip “LL” field (C) and Tongue Body “TB” field (D) stabilized with no activation peak. Their low activation values reflect the cross-field inhibition of each field by the Tongue Tip “TT” field.

Representing each articulator as its own field in the model with Voicing as one separate field reflects the purpose of a planning field, which is to compute a single production value (or set of values in a multi-dimensional field) based on one or more (potentially conflicting) inputs. The present model assumes that these planning fields are the mechanism by which phonological planning of any utterance is achieved, that

is, they are not specific to this experimental task. The design of the planning fields should therefore reflect the general demands of speech production. The example syllables relevant to the task modeled here are very simple, in that the initial consonants of each syllable require the specification of only one articulator, its constriction location, and voicing value. However, one does not have to look beyond the inventory of English consonants to find examples where more than one articulator from the present set would need to be specified simultaneously: e.g., /w/ requires labial and dorsal gestures; /ɹ/ requires labial, coronal, and dorsal gestures, etc. A single DFT planning field cannot generate more than one “winning” value. If, say, a Place field were designed where the parameter axis were all of the potential constriction locations in the vocal tract were specified, the mechanisms of DFT would not be able to generate output specifying two constrictions, e.g., one at the lips and one at the velum. Voicing, on the other hand, lends itself naturally to being defined on one field, since it is not logically possible to implement one CV utterance that has two VOTs. There can be only one value of the relative timing between two gestures. In the model presented here, VOT is independent of articulator. Phonetic studies have shown that there is a relationship between articulator and VOT, though this relationship is complex (Cho & Ladefoged, 1999). This aspect of the model design may therefore be a simplification to a degree, which is discussed further in section 7.3.4 of Chapter 7.

With these considerations in mind, the design of the model presented here corresponds closely with the parameters used in Articulatory Phonology (Browman &

Goldstein, 1986, et seq.) and is developed using the parameters of that framework (for a similar implementation, see Kirov & Gafos, 2007), though the model could be applied to any appropriate system of representation. In the present model, there is a field for each “tract variable”<sup>1</sup> in Articulatory Phonology. Each tract variable parameter that needs to be set is defined as an axis in that field. Therefore, even though the examples here show only one parameter axis (constriction location) for each articulator, a fully-developed model would include another parameter axis for constriction degree. Since constriction degree was constant in all experimental conditions being considered, it is not included in the model for computational and expositional simplicity.

The three articulator fields are coupled by fully-symmetric cross-field inhibitory links: when any activation level of one field crosses some threshold ( $\chi$ ), that field inhibits equally the other two fields by subtracting a constant amount of activation ( $q$ ) from every point of the other fields. There is no inhibition of the articulator fields by the VOT field, nor of the VOT field by any of the articulator fields. Cross-field inhibition is commonly used in other DFT models (e.g., Hock, Schöner, & Giese, 2003), as well as other interactive-activation models, e.g., the

---

<sup>1</sup> The design of the articulator fields in the present model differs slightly from Articulatory Phonology in that constriction degree and constriction location of the same primary oral articulator are treated as separate tract variables in Articulatory Phonology (Browman & Goldstein, 1989), while here they are treated as two parameter axis of the same field. Nothing crucial in the present model hinges on this difference, and the present model design could be modified to reflect Articulatory Phonology more closely. Whether this change is warranted is a question for further theoretical and empirical investigation.

TRACE model of speech perception (McClelland & Elman, 1986) and the production models of Dell (1986) and (Meyer & Gordon, 1985). The details of the cross-field inhibition are spelled out below in section 6.3. “Model Dynamics”.

The planning fields evolve over time as a result of combinations of different inputs, based on the dynamics of the model. These inputs are described and defined in the next section.

### 6.2.2. Input

There are three sources of input to the evolving planning fields. One input source corresponds to the parameter values for the required response, and one corresponds to the parameter values for the auditory distractor perceived during the planning of the utterance to be produced. The last source of input reflects the subject’s expectations of possible responses on a given trial. All inputs are represented as two-dimensional distributions of activation levels across the spectrum of possible values for a given parameter (see Chapter 5, section 5.3.2.4 for more detail). Although not required by the framework, all inputs in the present model are normal distributions with a mean ( $val$ ) and standard deviation ( $\sigma$ ) corresponding to reasonable values for the particular input, defined in (7).

$$(7) \quad \text{activation}_{\text{input}} = e^{-(x-val+noise_i)^2} / 2\sigma^2$$

The input for a required response (“response” in Figure 29) is assumed to be retrieved from the speaker’s representation of the parameter values required for the

response indicated by the visual cue in the experimental task. For example, the input for a required response of *ta* results in one input to the TT field with  $val = 0$  (corresponding to a constriction location of alveolar) and one to the Voicing field (Figure 30) with  $val = 55$  ms. The mean of the input distribution for the same utterance varies slightly on each trial due to random noise ( $noise_i$ ) included for both the articulator constriction location and VOT values. This noise term is a number drawn from a uniform distribution of VOT values between  $-27.5$  and  $27.5$  ms VOT to  $val$  on each trial. This variation reflects the fact that those values will vary for a given speaker across utterances due to factors unrelated to the factors included in the model. A simplification in the response input is that there are no differences in VOT input distribution value among voiceless responses beyond the noise added to the input, though VOT of English voiceless stops have been shown to vary by place of articulation (Lisker & Abramson, 1964). This assumption seems reasonable given the data from experiment 1. VOT measurements were made of all responses in both blocks, and no significant differences were found between *ta* and *ka* (see section 7.3.3 of Chapter 7).

The inputs corresponding to the auditory distractor are the parameter values of the stimulus that were perceived during the trial (“distractor” in Figure 29). The means and standard deviations for the distractor input distributions are the same as for the comparable response inputs, though there is no noise added to the mean of the distractor input, since the sound file played for a given distractor (e.g., *da*) never

changed within or across experiments. Some assumptions were made about the distractor input. First and foremost, this perceived stimulus is assumed to serve obligatorily and involuntarily as input to the ongoing planning in the production process. The model relies crucially on this link to account for the observed differences in RTs reported in Chapter 3, and in other studies. Second, it is assumed in the model that the speaker perceived the distractor unambiguously, never mistaking the place and/or voicing of the distractor stimulus. This means that for a trial where the distractor was *da*, there was input to the TT field, but none to the LL or TB fields, and the input to the Voicing field was always the same voiced VOT value. This is a simplification, as there is good evidence that hearers do confuse certain places of articulation and voicing, and that these sources of confusion can interact (Miller & Nicely, 1955). However, in the present experimental task, only two linguistic distractors were heard within a block of 280 trials. Given the amount of repetition of two sound files in a block (one for each distractor), it seems reasonable to expect that subjects were attuned the properties of the two syllables.

The last sources of input to the planning fields are called “pre-shapes”. Pre-shapes are raised activation levels that reflect the subject’s reasonable expectation about what response was possible on a given trial (“Task knowledge” in Figure 29). These raised activation levels are not sufficient to generate a response without other input, reflecting the fact that subjects did not reply until cued to do so. Other models in the DFT literature (e.g., Erlhagen & Schöner, 2002; Thelen, Schöner, Scheier, &

Smith, 2001) refer to stimulus-induced types of input—response and distractor input in the present model—as “specific inputs”, while the pre-shape input would be classified as “task input”, since it was task-specific rather than stimulus-induced. These pre-shape inputs are implemented the same way in all blocks of both experiments. In experiment 1, the articulator was always known but there were two possible values for voicing with 50% probability for either on a given trial. In simulations of this experiment only the articulator field corresponding to the known response have pre-shape input, defined by (7) with the same appropriate mean and standard deviation used for the response and distractor inputs. On the other hand, the pre-shape for the Voicing field is the sum of two distributions, again each defined by (7), with one having the appropriate mean and standard deviation for a voiced response and the other for a voiceless response. In experiment 2, the voicing was always known but there were two possible response articulators on a given trial within a block. In the simulation of this experiment there is input to three fields: one input to the articulator field corresponding to one of the possible responses on that trial, one input to the articulator field corresponding to the other possible response, and one input to the Voicing field. The pre-shapes interact with other effects on RTs in the model, and section 6.4 will show that the pre-shapes account for another aspect of the RT results in Chapter 3 that otherwise lacks explanation.

The three inputs differ in how long they persist as input to the evolving fields, and in their strength relative to one another. The specific values are given in section 6.5 and discussed in section 6.6.1.

### 6.2.3. Monitor and Implementation

The Monitor component determines when activation has built up in required fields to a level that is sufficient to send to Implementation, based on a criterion value ( $\kappa$ ). The criterion value  $\kappa$  is the same across all four planning fields. The decision criteria for the Monitor are straightforward. The Monitor waits until the activation level for some  $x$  value in both the Voicing field and one articulator field reach criterion. At that point it chooses the parameter values from those two fields with the highest activation level to be sent to Implementation. This has the practical effect that whichever field evolves more slowly determines the RT on the trial.

The Implementation component effects the motor control of the articulators needed to produce the planned utterance, and is not actually included in the simulations run below. This motor control system could be, e.g., either the Task Dynamics Model of Saltzman and Munhall (1989) or Guenther's DIVA model (Guenther, 1995; Guenther, Ghosh, & Tourville, 2006).

## 6.3. Model dynamics

The dynamics of each of the three articulator planning fields (LL, TT, and TB) are controlled by the equation in (8). The interaction term,  $interaction(x, t)$ , the “engine”

that drives the evolution of the activation field through local excitation and global inhibition, is defined in (9). The interaction term induces changes in the field as some value(s) of  $x$  approach a “soft” threshold ( $\theta$ ), which is determined by a sigmoid threshold function, defined in (10).

$$(8) \quad \tau dA(x, t) = -A(x, t) + h + p(\text{input}_{\text{PRESHAPE}}(x, t)) + r(\text{input}_{\text{RESPONSE}}(x, t)) + d_{\text{artic}}(\text{input}_{\text{DISTRCTOR}}(x, t)) - \text{inhibition}_{\text{CROSS-FIELD}}(x, t) + \text{interaction}(x, t) + \text{noise}$$

$$(9) \quad w(x) = w_{\text{excite}} e^{-(x^2/2\sigma_w^2)} - w_{\text{inhibit}}$$

$$(10) \quad f(u) = \frac{1}{1+\exp[-\beta(u-\theta)]}$$

These equations, which are the core of the DFT framework, are explicated in section 5.3 of Chapter 5. In summary,  $dA(x, t)$  is the change in activation level  $A$  of  $x$  at time step  $t$ . The rate of evolution of the field is controlled by  $\tau$ , with larger values of  $\tau$  resulting in slower evolution of the field.  $h$  is the resting level of the field. The inputs defined in the previous section are added to the field, when appropriate, by the terms  $\text{input}_{\text{PreShape}}(x, t)$  and  $\text{input}_{\text{Response}}(x, t)$  and  $\text{input}_{\text{Distractor}}(x, t)$ . The variables<sup>2</sup>  $p$ ,  $r$ , and  $d_{\text{artic}}$  encoded the relative strengths of the inputs. The cross-field inhibition introduced by any other articulator field(s) above threshold ( $\chi$ ) is added by the term  $\text{inhibition}_{\text{CROSS-FIELD}}(x, t)$ . The use of a soft threshold ( $\theta$ ) means that some  $x$  values below  $\theta$  do engage the interaction term, but the contribution to the interaction of those

---

<sup>2</sup> The word “variable”, rather than “parameter”, is used to describe terms in the model and DFT equations that have constant values in the model (e.g.,  $p$ ,  $\tau$ ,  $\chi$ , etc.). This is to avoid confusion with phonological properties (e.g., VOT, constriction location), which are referred to throughout as “parameters”.

activation values less than  $\theta$  diminishes with distance from  $\theta$  (see section 5.3.2.5 in Chapter 5 for more details). The system is therefore non-linear due to this soft threshold, in that incremental changes in activation levels have a non-uniform effect on its evolution. Noise is added to introduce stochastic behavior into the model evolutions.

The dynamics of the Voicing field are controlled by the equation in (11). The equation that controlled the dynamics of the Voicing field evolution differ from the

$$(11) \quad \tau dA(x, t) = -A(x, t) + h + p(\text{input}_{\text{PRESHAPE}}(x, t)) + r(\text{input}_{\text{RESPONSE}}(x, t)) + d_{\text{voice}}(\text{input}_{\text{DISTRACTOR}}(x, t)) + \text{interaction}(x, t) + \text{noise}$$

one that governed the articulator fields in only two regards. One is that there is a separate weighting variable  $d_{\text{voice}}$  for the distractor voicing value, i.e., it was not the same factor as the distractor weighting factor of articulator ( $d_{\text{artic}}$ ). The reason for this difference has to do with the different dynamics at play in within-field vs. cross-field inhibition in the model, which is discussed in detail in section 6.6.1. The other is that there is no cross-field inhibition term included in the dynamics, because the evolution of the articulator planning fields do not interact with the evolution of the Voicing field.

The rest of this section illustrates how the inputs and dynamics combine to yield the RT modulations observed in various experimental conditions. The particular variable values are given in section 6.5, then discussed in detail in section 6.6.1.

### 6.3.1. Pre-shapes

“Pre-shapes” are a source of input to the planning fields that reflect priming based on the subject’s reasonable expectations about possible responses on a given trial. Pre-shapes from the two experiments from Chapter 3 are shown in Figure 31, which show the evolution of the activation fields with no other input, i.e., no cued response and no distractor. The activation level of the LL field in Figure 31A and C provides a baseline of the resting level of a planning field with no pre-shape, since in neither block was the potential response a labial.

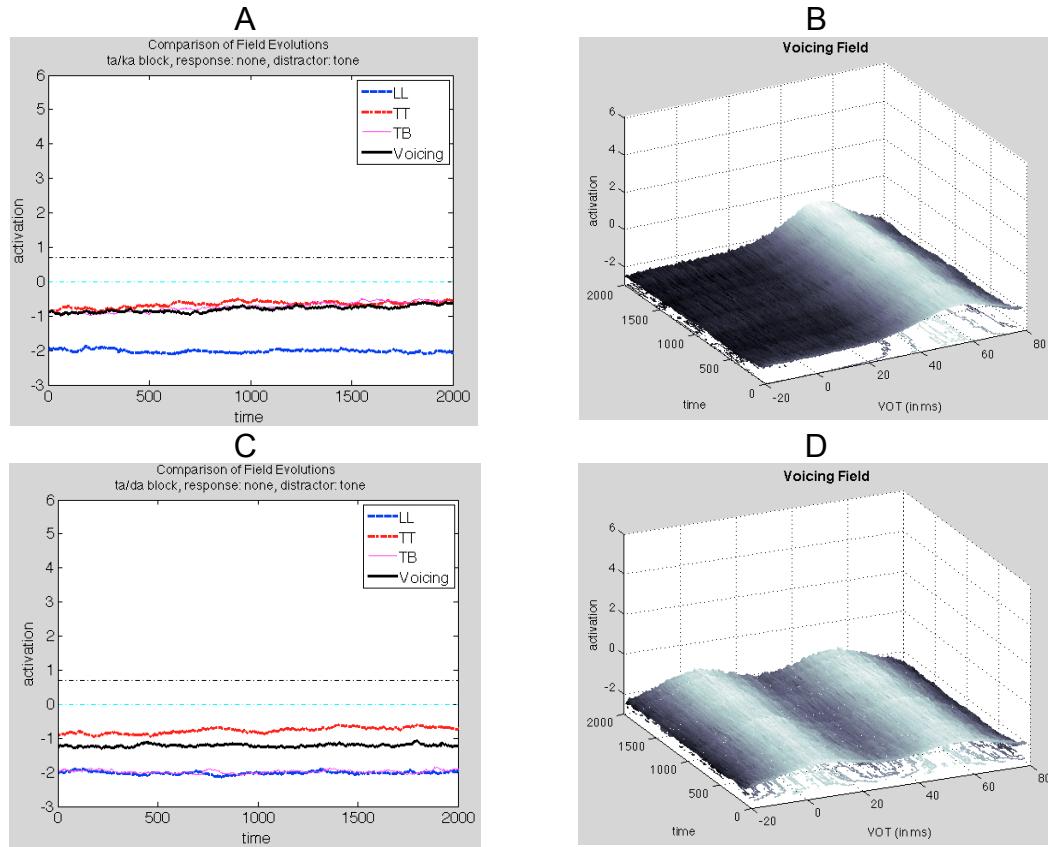
The same amount of pre-shape input ( $p$ ) is added to the appropriate fields to raise the activation level of the parameter values of any potential response. For example, Figure 31A shows the pre-shape input for a block in experiment 2 where the only two possible responses are *ta* and *ka*. The subject knows before the trial starts that the response will involve either the tongue tip or tongue back. Therefore in this block, there is pre-shape input to the TT field and to the TB field, each of the same amount ( $p$ ). The subject also knows that the response will be voiceless. Therefore, there is pre-shape input to the Voicing field for the voiceless VOT value (~55 ms), also in the amount  $n$ .<sup>3</sup> The result is that the levels for all 3 of these fields (TT, TB, and Voicing) are higher than for the LL field. Figure 31B shows the pre-shape of the Voicing field in the same simulated trial from experiment 2, which has a small persistent peak of activation of the voiceless VOT values. The pre-shape of the TT and

---

<sup>3</sup> Inputs are not doubled when the probability is 100% instead of 50%, as this would eventually lead to some field value reaching criterion.

TB fields are similar to the Voicing field, reaching a similar maximum activation, just greater than  $-1$ .

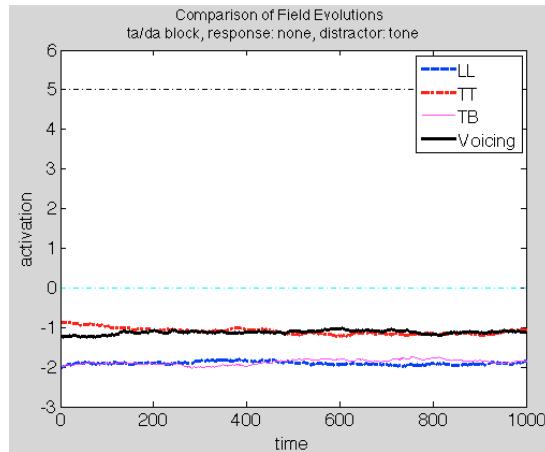
The pre-shapes for experiment 1 behave differently, as can be seen by comparing Figure 31C/D with A/D. In the block shown for this simulated trial, the possible responses are *ta* or *da*. Pre-shape input for the TT field is therefore the same



**Figure 31. Pre-shapes without other input.** (A) compares the maximum activation for the pre-shapes of the four activation fields in a block from experiment 2 where the potential responses are *ta* or *ka*, with (B) showing the pre-shape of the Voicing field on the same simulated trial. (C) compares the maximum activation for the pre-shapes of the four activation fields in a block from experiment 1 where the potential responses are *ta* or *da*, with (D) showing the pre-shape of the Voicing field on the same simulated trial. In (A) and (C), the black dash-dotted line shows the within-field interaction term “soft” threshold ( $\theta$ ) at activation level = 0.7, and the cross-field inhibition threshold ( $\chi$ ) at activation level = 0.

as in the example from experiment 2 above, in the amount  $p$ , and the field maintains a pre-shape activation level of just greater than  $-1$ . However, since the response could be either voiced or voiceless, there are two inputs to the Voicing field, both also of amount  $p$ , one corresponding to a voiceless response and another corresponding to a voiced response with a VOT of about 5 ms, (Figure 31D). The introduction of two inputs to the same field that are sufficiently far away from each other results in the interaction term engaging. This results in a lowering of the maximum activation levels in the Voicing field, due to the lateral inhibition introduced by the two incongruent inputs.

It may not seem from Figure 31C that the interaction term should have been active in the presence of only pre-shape input, since the maximum value of the Voicing field never seems to get sufficiently close to the interaction term threshold  $\theta$  (0.7, indicated in Figure 31A/C as the black dash-dotted line) to induce within-field inhibition. Recall, however, that  $\theta$  is a soft threshold, and that activation levels below  $\theta$  do engage the interaction term as determined by the sigmoid function defined in (10) and explained in more detail in section 5.3.2.5 in Chapter 5. To demonstrate that the lower level of activation for the Voicing field is due to the within-field inhibition introduced by the interaction term, the same trial shown in Figure 31C can be simulated again, though this time with  $\theta = 5$ , which is sufficiently high that no  $x$ -value activation level gets close enough to the soft threshold to engage the interaction term. No within-field inhibition should be introduced.



**Figure 32. Evolution of the Voicing field for a simulated trial with no within-field inhibition. The within-field interaction term threshold  $\theta$  (dash-dotted black line) is raised on this simulation to 5 (instead of .7 in Figure 31), resulting in no within-field inhibition introduced by the incompatible VOT values.**

The comparison of the resulting activation levels are shown in Figure 32, which shows that in this simulation, the Voicing field pre-shape does indeed settle at the same level as the TT field. The only difference between the simulation in Figure 32 and the simulation in Figure 31C is the change in  $\theta$ . Another simulation of the same trial (not shown) where  $\theta$  is left at 0.7 but the amount of cross-field inhibition is 0 results in the same difference between the Voicing and TT fields as shown in Figure 31C. These two simulations demonstrate clearly that the interaction term is the cause of the difference between the Voicing pre-shapes in experiment 1 (Figure 31C) and experiment 2 (Figure 31A).

Lastly, Figure 31 shows that the amount of activation increase introduced by the pre-shapes is not enough to cause the activation of any field value to rise sufficiently to reach criterion via the DFT field-internal dynamics (as does Figure 32).

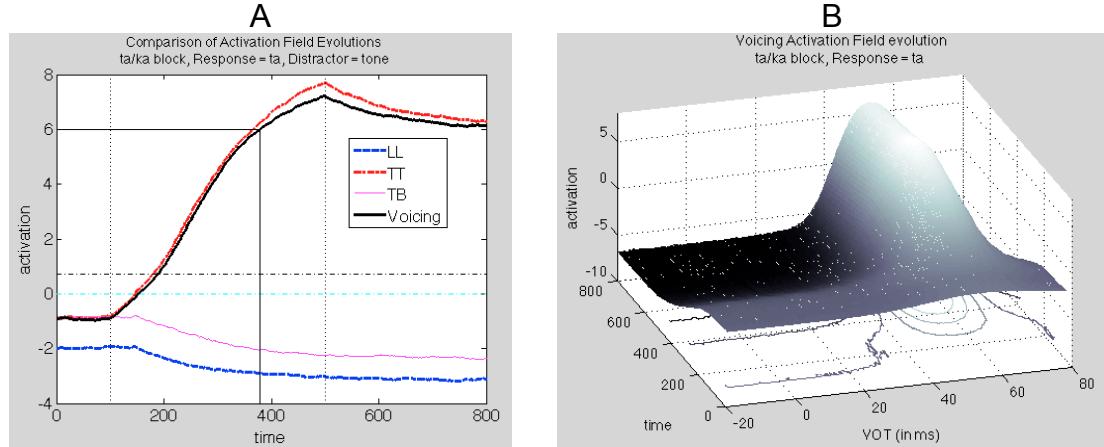
The evolutions in Figure 31 evolve for 2000 time steps—much longer than is required to simulate other trials in the examples below—and stays at roughly the same level throughout the evolution.

### 6.3.2. No inhibition: Tone condition

Trials where there is no linguistic distractor represent the simple case where a subject sees a cue indicating what response is required (see discussion, Chapter 4, section 4.2.2), retrieves the appropriate representations for the necessary parameter values for that response, and executes the required phonological planning. This case serves here as a baseline reference for the other experimental conditions where linguistic distractors are introduced. An illustration of the dynamics of this simplest case is illustrated in Figure 33.

The example trial is of a *ta-ka* block experiment 2, where subjects know the voicing of the response (voiceless) but not the articulator, which has a 50% probability of being TT (for *ta*) and a 50% probability of being TB (for *ka*). The first 100 time steps show the state of the activation fields at the beginning of a trial, reflecting the presence of no input other than pre-shape information. The activation level for the LL field is at the normal resting level of a field in the system (about -2), since no potential response involves the Lower Lip. The trial-initial activation level for the Voicing field is higher than the level of the LL field, since the voicing is known. The trial-initial

levels for the TT and TB fields are also higher than for the LL field in anticipation of either a tongue tip or tongue back production.



**Figure 33. Activation field evolutions in the Tone condition of experiment 2. (A)** The evolution of the maximum activation levels of the four planning fields is shown for one simulated trial of the response-distractor task when there is no linguistic distractor (i.e., the Tone condition). The possible responses are *ta* or *ka*, with each having 50% probability. The required response for this trial is *ta*. Vertical dotted lines indicate the time during the evolution when the planning field has input based on the parameter values of the required response, i.e., 400 time steps from 100 to 500. The dash-dotted cyan line at activation level = 0 shows the cross-field inhibition threshold  $\chi$ . The dash-dotted black line at 0.7 shows the within-field interaction term threshold  $\theta$ . The horizontal line at activation level = 6 represents the criterion value  $\kappa$  used by the Monitor function. On this trial, the maximum activation level of the TT field (dash-dotted red line) crosses criterion second and therefore determines the RT, indicated by the vertical solid black line. **(B)** Evolution of the Voicing field for the same trial. The evolution of the TT field in this trial is qualitatively the same as the Voicing field, the only difference being the units of the parameter axis.

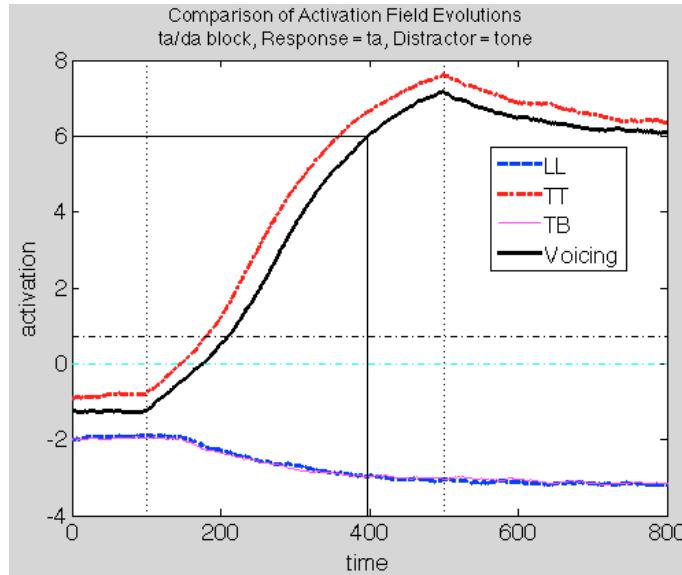
After 100 time steps, the fields starts to evolve based on the input of the required response to the Voicing and TT planning fields, as evidenced by their steady rise above their trial-initial states. For about 60 time steps the changes in the activation level are due largely to the reinforcing effect of this input and the pre-shape. No cross-

field inhibition is introduced, as evidenced in Figure 33A by the activation levels for the LL and TB fields not materially changing from their trial-initial states.

About 60 timesteps after the introduction of the response input (i.e., just after time step 160), the maximum activation of the TT field breaches the cross-field inhibition threshold  $\chi$ . This has the effect of subtracting a constant amount of activation overall from the other two articulator fields at every time step in the evolution of the field. The maximum levels of the LL and TB fields can be seen to drop, each by the same absolute amount, starting at the same time that the TT field maximum breaches  $\chi$ . These fields ultimately stabilize at a level which is the same as their trial-initial level (about -2), minus the amount of the cross-field inhibition (1.75). Though the interaction term had been contributing some local excitation and lateral inhibition as the activation levels approached  $\theta$ , at just about time step 200, both the TT and Voicing reach the within-field interaction term threshold  $\theta$ . The influence of the interaction term action become more visible. First, the effect of local excitation can be seen in Figure 33A as the increase in the rate of activation level rise for the Voicing and TT fields after the within-field threshold is crossed—that is, the slope of the TT and Voicing lines is greater after time step 200 than it is between time steps 100 and 200. Second, Figure 33B shows the effect of global inhibition after time step 200. As the VOT activation levels around 55 ms rise above the within-field threshold, VOT activation levels that are far away from 55 ms are lower than the trial-initial level (around -7). These lower activation levels are a result of the maximum global

inhibition introduced by the interaction term from the VOT values above the within-field threshold.

This pattern continues until the Monitor function determines that some  $x$  value activation level of both the Voicing field and one articulator field have both reached criterion  $\kappa$ , at which point the parameter value from each field whose activation level has reached criterion is sent to implementation. In the example trial shown in Figure 33, the maximum activation for the Voicing field reaches criterion first, quickly followed by the TT field. Therefore, on this particular trial, the time step when the TT field activation reaches criterion was logged as the “RT” for this trial, indicated by the vertical straight black line in Figure 33A at time step 379. This equates to a RT of 279 since the start of input to the field for the response roughly indexes the presentation of the visual cue started at time step 100. This is somewhat of a simplification, since this timing does not account for visual processing, retrieval of other representations from long-term memory, and other requirements demanded of the task. For any trial with these potential responses and no linguistic distractor, sometimes it is the TT field that crosses second and therefore determines the RT of the trial, due to the stochasticity introduced by the noise in the evolution, see (7).



**Figure 34. Activation field evolutions in the Tone condition of experiment 1.**

Figure 34 shows the evolution of the activation field levels for the Tone condition in experiment 1, where the articulator of the response is known but the voicing is not. There is one slight difference between this trial and the one depicted in Figure 33. The pre-shape of a trial in experiment 1 differs from the pre-shape of a trial in experiment 2 in that in the former, the activation level of the Voicing field is lower than the activation level of the TT field due to the effects of within-field inhibition in the VOT field arising from the incompatible pre-shape inputs in experiment 1, as explained in section 6.3.1. The Tone condition in this experiment also represents the evolution of the fields with no distractor inhibition, but due to the fact that the Voicing field starts out lower than the TT field, the Monitor always waits for the Voicing field to reach criterion to determine the RT in trials of this type. This is unlike the

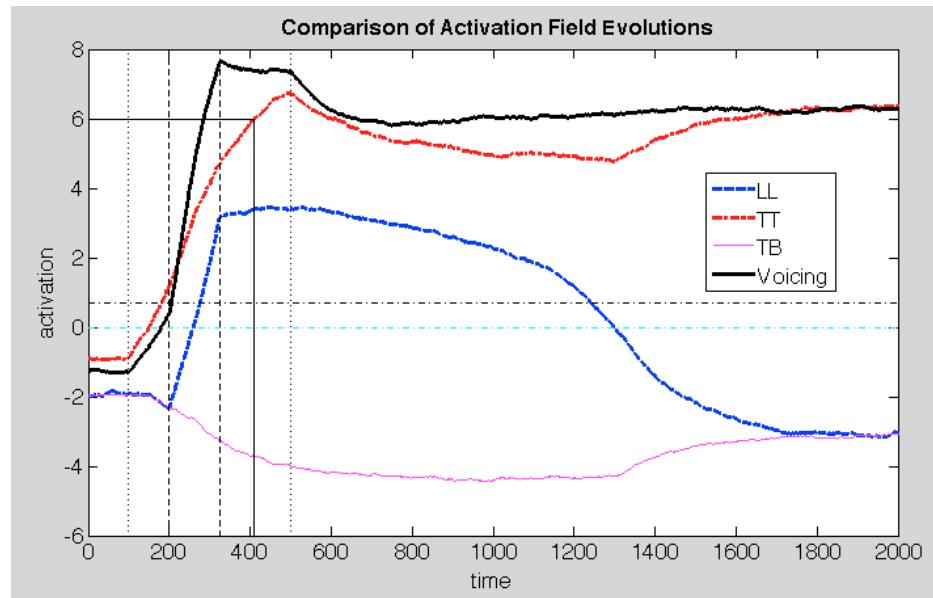
experiment 2 trial, where the RT could be determined by either field due to the similar starting point of the rate of rise of the TT and Voicing fields.

### 6.3.3. Cross-field inhibition: Congruent condition, Experiment 1

This section shows how the behavior of the model changes from the simple Tone condition above when cross-field inhibition is introduced. This is exemplified by a trial in the Congruent condition of experiment 1. In this trial, the two possible responses are *ta* or *da*, and the subject is cued to respond *ta*. The distractor played is *pa*, therefore matching the response in voicing but differing in articulator. The Stimulus Onset Asynchrony (SOA) between the cue indicating the required response and the distractor is 100 time steps on this trial. The evolution of the fields for one trial is shown in Figure 35.

Figure 35 shows that on this trial the Monitor determines the response values at time step 419 (equivalent to a RT of 319) which is slower than the RTs shown in Figure 33 and Figure 34. This section explains how cross-field inhibition introduced by the distractor accounts for the difference in RT between this trial and the one illustrated in Figure 34. The most important difference between this trial and the trials illustrated in Figure 33 and Figure 34 is the presence of input to the fields from a distractor, which is indicated in Figure 35 by the vertical dashed lines. Since the SOA for this trial is 100 time steps, the distractor input starts at time step 200, 100 steps after the start of the input for the required response. The evolution of the fields

progresses effectively the same as in Figure 34 from the start of the response input at time step 100 until the introduction of the distractor input at time step 200. At this point several changes can be seen. First, the rate of activation level rise for the Voicing field increases notably because the response and the distractor have the same values for VOT. The reinforcing VOT inputs of the response and the distractor combined to increase the amount of activation added on each time step for the voiceless response. The interaction term kicks in maximally when the Voicing field crosses the within-field threshold roughly at time step 200, resulting in even greater activation increases for the input VOT values than the sum of the input values, due to local excitation and no inhibition of either input by the other. This is seen by the steep slope of the Voicing activation line in Figure 35, which crosses criterion ( $\kappa$ ) at about time step 300. This is earlier than the trial in Figure 34, where the Voicing activation crosses criterion at time step 379. Functionally, this means that on trials where the response and distractor share voicing, the RT is never determined by the Voicing field, which always reaches criterion earlier. RT is determined by the evolution of the articulator field, which in this simulated trial is the TT field.



**Figure 35. Comparison of activation field evolutions showing cross-field inhibition.** A simulated trial from the Congruent condition of the *ta-da* block of experiment 1, where the required response is *ta* and the distractor is *pa*. This figure is comparable to Figure 34, but with the addition of a distractor. The duration of the input from the distractor (125 time steps) is indicated by the vertical dashed lines. The distractor input is introduced 100 time steps after the start of the response input (SOA = 100 time steps).

Second, at the time step where the distractor input starts (200), the activation level of the LL field begins to rise due to the *pa* distractor, in parallel with but lower than the activation level of the TT field. Throughout the duration of the distractor input, the LL activation level continues to rise, despite the cross-field inhibition being exerted on it by the TT field, which has already breached the cross-field threshold by the time the distractor input starts. At about time step 300, the activation level of the LL field also breaches the cross-field threshold and starts to inhibit the TT field (and the TB field). This inhibition has the effect of slowing the rate of rise of the TT field the entire time until it reaches criterion. Since it is always the articulator field on these

trials that determines the RT, the slow down of RTs in the congruent condition of experiment 1 compared to the Tone condition can be attributed unambiguously to the effect of cross-field inhibition introduced by the articulator of the distractor.

Other effects of cross-field inhibition are visible in Figure 35, though these effects do not have any bearing on the RT differences being accounted for in this task. They are presented and discussed here solely to further illustrate the workings of the cross-field inhibition. The activation level of the LL field, though never reaching criterion, does get sufficiently high that it crosses the within-field interaction term threshold. The result is that even when the input from the distractor has stopped, the activation level of the LL field is sufficiently high that the inherent dynamics of the interaction term in the LL field keep the activation peak well above the cross-field inhibition threshold for a considerable amount of time. The LL field therefore continues to inhibit the TT and TB fields until just after time step 1200. The level of the TT field, even though it has passed criterion, stays below the activation level of the Voicing field while the LL level is relatively high. However, the LL field is also being inhibited by the TT field during this time, and eventually this inhibition overcomes the influence of the DFT interaction term, suppressing the LL level below the point where the field-internal dynamics can maintain the peak in activation. Once the LL field activation level drops below the cross-field inhibition threshold at about time step 1230, the activation level of the TT field starts to rise toward the uninhibited stabilization level just above 6, where it ultimately stabilizes, similar to the Voicing

field. The TB activation level also starts to rise toward a stable state that includes only the inhibition introduced by the TT field, which is where it converges with the LL field level.

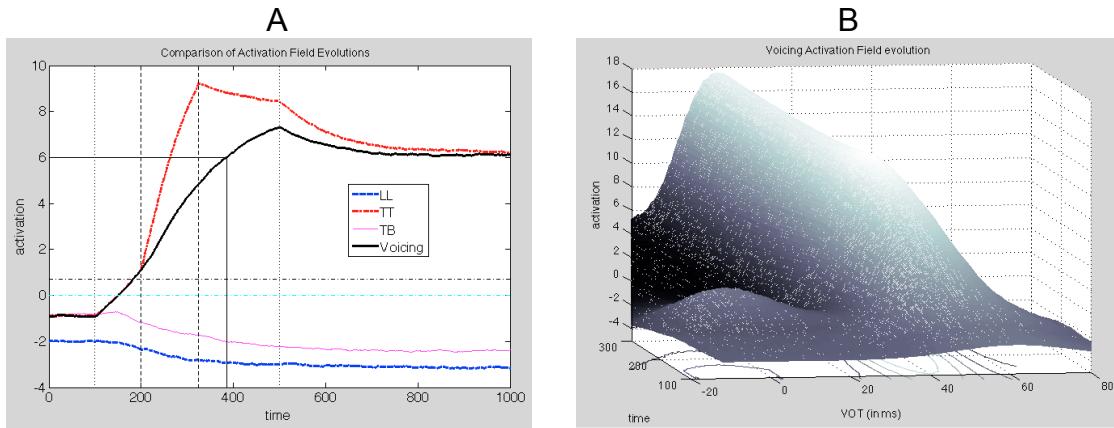
#### **6.3.4. Within-field inhibition: Congruent condition, Experiment 2**

This section shows how the behavior of the model changes from the simple Tone condition above when within-field inhibition is introduced, rather than cross-field inhibition as in the case in the preceding section. This is exemplified by a trial in the Congruent condition of experiment 2. In this trial, the two possible responses are *ta* or *ka*, and the subject is cued to response *ta*. The distractor played is *da*, therefore matching the response in articulator but differing in voicing. The evolution of the fields for one trial is shown in Figure 36.

On the example trial shown in Figure 36, the Monitor determines the response values at time step 391 (equivalent to a RT of 291), slower than the RT shown in Figure 33. The trial shown in Figure 36 is comparable to the trial shown in Figure 35 in that the SOA is 100 time steps, and the durations of the inputs to the fields are the same, as were all threshold values. The pre-shapes in Figure 36A, although they do not qualitatively influence the effect of congruency in simulations of this experiment, are different than those in Figure 35. The trial-initial pre-shape activation level of the TT, TB, and Voicing fields are all the same, at around  $-1$ , while the trial-initial pre-shape activation levels of the LL field is at  $-2$ . For the trial shown in Figure 36, the

activation levels of the TT and Voicing fields begin to increase when input for the required response is introduced at time step 100. The levels of the TB and LL fields remain relatively unchanged for about 50 time steps, but at approximately time step 150, the TT field breaches the cross-field threshold. From this point, the levels of the TB and LL fields begin to drop due to the cross-field inhibition introduced by the TT field. The TT and Voicing fields continue to rise at effectively the same rate until the input from the distractor is introduced at time step 200. Since the articulator input of the distractor is qualitatively the same as the input for the response, i.e., an alveolar constriction location in the TT field, the activation level of the TT field begins to rise much more sharply due to the combined effects of two reinforcing inputs and the effects of the interaction term, analogous to the behavior of the Voicing field at the comparable point in the trial shown in Figure 35.

Just before time step 200, the Voicing field also crosses the within-field threshold, increasing the influence of the interaction term. The effects of this can be seen in Figure 36A as a slight increase in the rate of rise of the activation level of the Voicing field just after time step 200. This increase in rate of rise is less than for the TT field in large part because the input to the TT is more than twice as strong as the input to the Voicing field, with the former having two compatible inputs instead of one, plus excitation from the interaction term. The rate of rise of the Voicing field of the trial depicted in Figure 36A is slower also than the Voicing field of the trial depicted in Figure 33A because the input of the distractor



**Figure 36. Comparison of activation field evolutions showing within-field evolution from a simulated trial from the Congruent condition of the *ta-ka* block of experiment 2, where the required response was *ta*, the distractor was *da*, and the SOA was 100 time steps. (A) compares the maximum activation levels of the four planning fields during the timecourse of the trial. (B) shows the evolution of the Voicing field.**

(voiced VOT) is incompatible with the input corresponding to the required response (voiceless VOT). This distractor input can be seen in Figure 36B. The distractor input is the smaller peak in activation to the left of the main activation peak building up as required for the voiceless response. In the trial depicted in Figure 36, the activation peak corresponding to the voiced distractor VOT input, despite being small and dropping off once the distractor input ends at time step 325, is sufficient to introduce enough within-field inhibition of the voiceless VOT values in the Voicing planning field to slow the rate of rise of that field compared to the Tone condition shown in Figure 33B, where there is no such distractor or within-field, lateral inhibition. After the peak of activation corresponding to the distractor VOT has gone, the voiceless VOT peak continues to rise based on the response input plus the action of the

interaction term. The Voicing field in trials for this condition therefore always reach criterion after the TT field, and is the source of the RT.

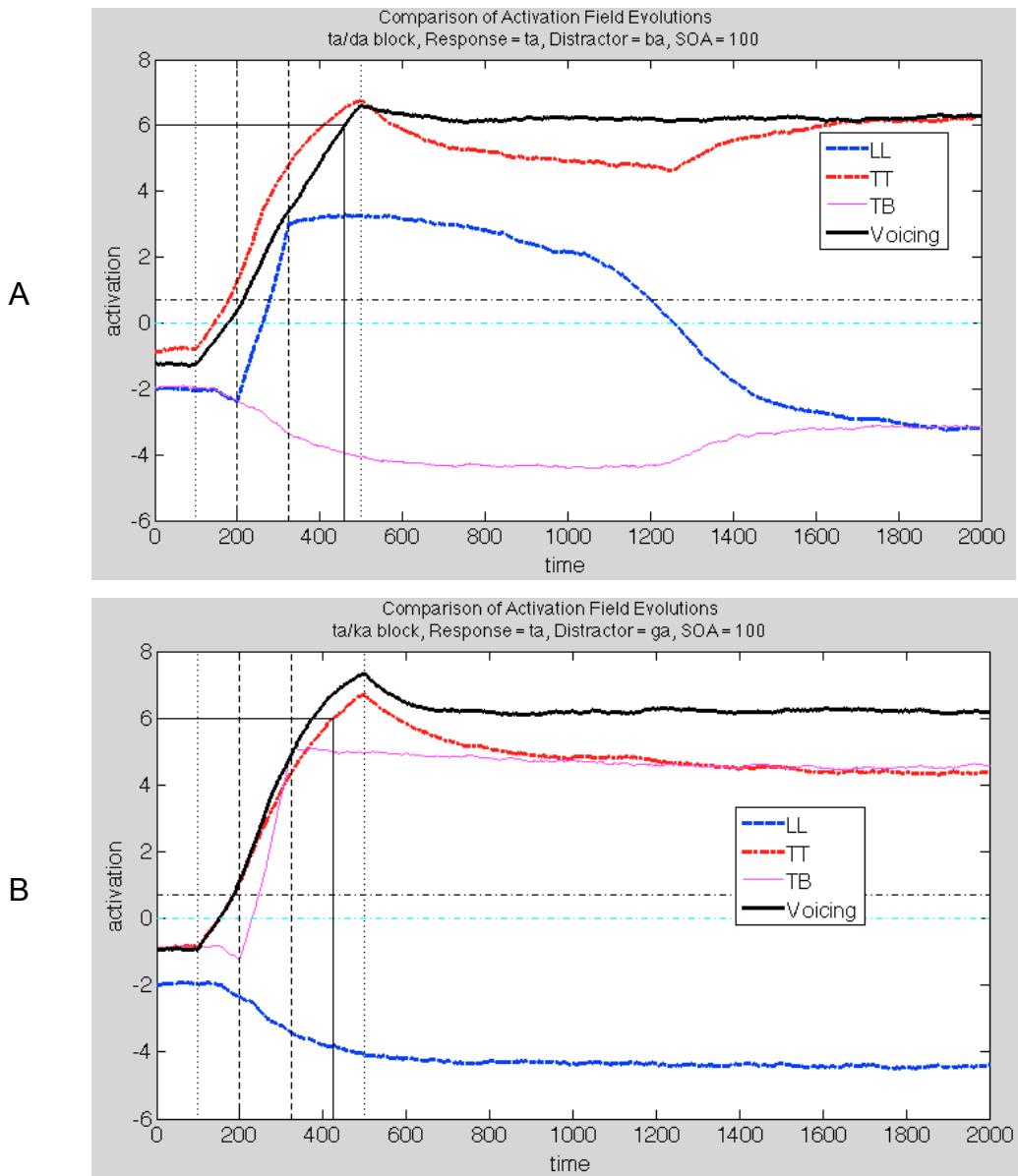
After the response input has stopped at time step 500, the activation levels of both the TT and Voicing fields drop off and stabilize just above 6, as they ultimately do in Figure 33 and Figure 35. Unlike the trial depicted in Figure 35, the TT field is not inhibited in the trial depicted in Figure 36 because the other two articulator fields remain below the cross-field inhibition threshold throughout the trial due to the early determination of the TT response as reinforced by the distractor.

### **6.3.5. Cross- and within-field inhibition: Incongruent condition**

The previous two sections have shown that while RTs in the Congruent condition in both experiment 1 and experiment 2 are longer than in the Tone condition, this difference is accounted for by two different sources of inhibition, depending on the experiment. In experiment 1, the slow-down vs. the Tone condition is attributable to cross-field inhibition introduced by a distractor mismatching the response in articulator, though matching in voicing. In experiment 2, the slow-down vs. the Tone condition is attributable to within-field inhibition introduced by a distractor mismatching the response in voicing, though matching in articulator. RTs in the Incongruent condition in both experiments—where the response and distractor mismatch in both articulator and voicing—are longer than RTs in the Congruent condition of the given experiment. This section demonstrates that while both within-

and cross-field inhibition are present in the evolution in the Incongruent condition of both experiments, the difference in RTs between the Congruent and Incongruent condition is attributable to whichever inhibition is not the cause of the RT slow down in the Congruent condition (vs. the Tone condition). That is, the difference between the Incongruent and Congruent conditions arises from the introduction of increased within-field inhibition in experiment 1, and from the introduction of cross-field inhibition in experiment 2.

Figure 37 depicts the evolution of the activation fields in a trial from the Incongruent conditions of experiments 1 and 2, where the required response is *ta* and the SOA is 100 time steps. Figure 37A shows a trial from the block in experiment 1 where the possible responses are *ta* or *da*. The source of the difference in RTs between the Congruent and Incongruent conditions can be seen by comparing the evolutions shown in Figure 35 and Figure 37A. The pattern of evolution of all of the articulator fields is qualitatively the same in both figures. The activation level of the TB field stays low throughout the evolution as neither the response or distractor involve the Tongue Back. The activation level of the LL fields rise due to the *pa* distractor enough to inhibit the TT field, but eventually drop back to resting level after the distractor input ends. The activation level of the TT field rises with the introduction of the *ta* response input, but is slowed (relative to the Tone case) by the cross-field inhibition introduced by the rise in the activation of the LL field. The activation level of the TT field reaches criterion at about the same time step, around 400, in both trials.



**Figure 37. Comparison of activation field evolutions in the Incongruent conditions of experiments 1 and 2. (A) A simulated trial from the Incongruent condition of the *ta-da* block of experiment 1, where the required response is *ta* and the distractor is *ba*. (B) A simulated trial from the Incongruent condition of the *ta-ka* block of experiment 2, where the required response is *ta* and the distractor is *ga*.**

The qualitative difference between the two trials is the evolution of the Voicing field. On the trial from the Congruent condition (Figure 35), the Voicing field rises very quickly due to the combined, congruent response and distractor inputs. The Voicing field therefore always crosses criterion much earlier than the TT field, so on Congruent trials the Monitor always determines the RT based on when the TT field reaches criterion. On the trial from the Incongruent condition (Figure 37A), the within-field inhibition due to the incompatible VOT value of the distractor input makes the rate of rise of the Voicing activation level slower than both the rate of rise of the TT field in the Congruent (Figure 37A) and the Voicing field in the Tone condition (Figure 34). In the Incongruent condition the Monitor determines the response values based on when the Voicing field reaches criterion, which is at time step 460 (an effective RT of 360), longer than in the Congruent condition shown (Figure 35).

Figure 37B shows a trial from the Incongruent condition of experiment 2 where the possible responses are *ta* or *ka*. A comparison of Figure 36A and Figure 37B of shows the source of the RT differences between the Incongruent and Congruent conditions in this experiment. In contrast to the comparison above for experiment 1, here it is the evolution of the Voicing field that is qualitatively the same because the distractor mismatches the response in VOT value in both conditions. In the Congruent condition it is this Voicing slow-down that determines the RT (which was 291), whereas in the Incongruent condition the TT field reaches criterion after the

Voicing field and therefore determines RT, which is at time step 420 (= RT of 320).

The cross-field inhibition introduced by the TB field of the distractor causes the rate of rise of the TT field in the Incongruent condition to be slower than in the Congruent or Tone (Figure 33A) conditions.

Another difference between the trials shown in Figure 37A and B is that in the trial from experiment 2 shown in B, the activation level for the distractor articulator (TB) reaches sufficient strength to stabilize just above 4 due to the field-internal interaction term dynamics, despite the cross-field inhibition being introduced in its evolution by the TT field. This stabilization does not occur in the trial from experiment 1 (Figure 37A). This also has the effect of lowering the level at which the TT ultimately stabilizes, compared to the level at which it stabilizes in experiment 1 (shown in Figure 37B). The difference in the two evolutions between the two Incongruent conditions is due to the difference in the pre-shapes in the two experiments. In experiment 2 the articulator of the distractor is one of the two possible responses on that trial, whereas in experiment 1 the articulator of the distractor is not one of the possible responses. The trial-initial level of activation for the distractor articulator is therefore higher in experiment 2 than in experiment 1. This higher start in the experiment 2 trial is enough to allow the activation level of the distractor's articulator (TB) field to reach a value sufficiently high that it can stabilize through the DFT-inherent mechanisms. The lower start in the experiment 1 trial

“disadvantages” the activation level of the LL field enough that it never reaches a value high enough to stabilize.

It seems tempting to interpret the differences in the stabilization of the distractor articulator fields as making some predictions, e.g., subjects at later RTs in the Incongruent condition in experiment 2 should seemingly either be at chance between implementing the articulator of the cued response (e.g., *ta*) or the distractor (e.g., *ka*), or (attempt to) produce a double articulation (e.g., *t*□*ka*), while in the case of experiment 1 they should reliably respond with the correct articulator (e.g., *ta*). Such conclusions are unfounded. The model as implemented does not address what happened to the fields after the Monitor chose production parameter values for implementation. It is common in many models of speech production (e.g., Dell & O'Seaghda, 1992; MacKay, 1987; Shattuck-Hufnagel, 1979) to have a process that actively resets activation levels of fields after a choice has been made, usually to avoid re-selection of the same field for a downstream production unit. This stabilization of the TB field activation level in experiment 2 persists in part because no such mechanism is included in the present model. The model does not currently address how the fields get reset after the Monitor makes its choice of production parameter values on a given trial, but it must be the case that some such function exists. In all of the conditions in the model, the activation levels of the production parameters reach a level where they stabilize, which by definition means that they do not return to resting (or trial-initial) state without the introduction of some inhibition from an outside

source (see section 5.4.3 in Chapter 5). The scope of the model as presented here is to account for the RTs observed in the experiments presented in Chapter 3 and in other studies. This requires only specifying the mechanism that determines the RT, which is what the Monitor does. The function of the Monitor could be expanded to introduce inhibition to all planning fields upon choosing its parameter values, in which case the equal stabilization of the TT and TB fields in the Incongruent case of experiment 2 would not be present. The evolutions are left without this function to illustrate the dynamical behavior of the fields. All activity after the Monitor selection point is of no consequence to the model as implemented and should not be used as a predictor of behavior at any time after that point.

The preceding sections have explained how the model accounts for the differences in RTs in the three conditions (Tone, Congruent, and Incongruent) of the two experiments presented in Chapter 3. The model also accounts for the Identity condition, which is not included in the present experiments but is reported in other studies (Galantucci et al., 2009; Kerzel & Bekkering, 2000) using the response-distractor task.

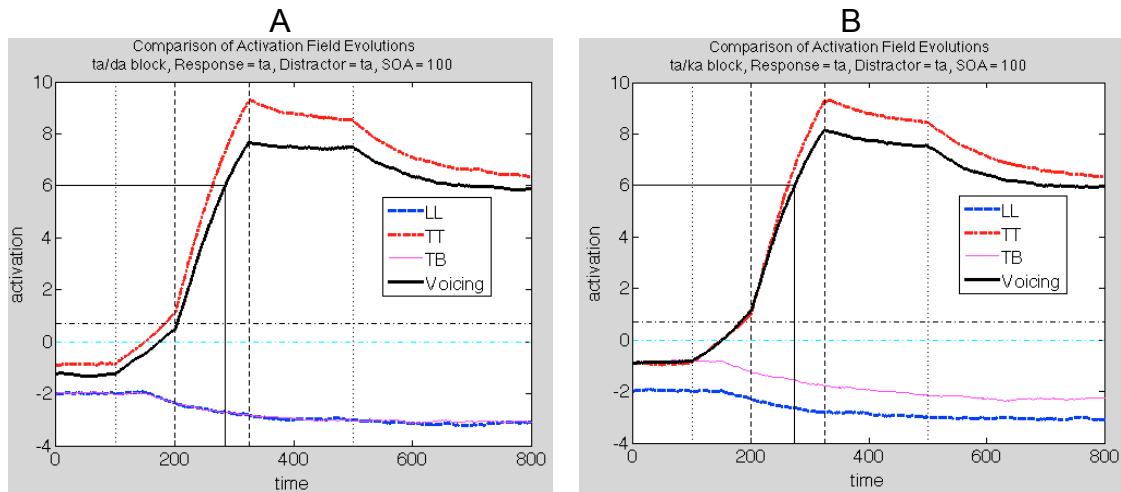
### **6.3.6. Reinforcing input: the Identity condition**

In all of the conditions described in the previous section, modulations of RTs are ultimately all attributable to the effects of inhibition—both within- and cross-field—introduced by incongruent parameter value input of a distractor. The effects of

congruent parameter values of a distractor on the evolution of a field's activation level are visible in the increased rate of rise in both of the Congruent cases, for the TT articulator field in experiment 1, and the Voicing field in experiment 2. However, these congruency effects have no material role in determining RTs. The effects of compatible input from the distractor are simply that the RT is always determined by the rate of rise of the planning field for the other, incompatible distractor input because the field for the compatible parameter reaches criterion very quickly. As outlined in Chapter 4, the results of the experiments in Chapter 3 plus the results of Galantucci et al. (2009) motivate the computational principle of excitation being necessary to account for the result that RTs in the Identity condition were shorter than the Tone condition (see also discussion in Chapter 3, section 3.4.2 justifying the direct comparison of the Galantucci et al. results with the results in Chapter 3). This section shows how the model derives the RTs in the Identity condition being shorter than in the Tone condition.

Figure 38 shows the evolution of two trials in a hypothetical Tone condition of both experiments from Chapter 3. As in the Congruent and Incongruent examples above, the response in both trials is *ta*, the SOA is 100 time steps, though in both of these examples, the distractor is also *ta*. Figure 38A shows the prediction of the model for the Identity condition in experiment 1, and Figure 38B shows the prediction for experiment 2. The Monitor determines the production parameter values at time step 284 (= RT of 184) for experiment 1, and at time step 274 (= RT of 174) for

experiment 2. These RTs are shorter than the RTs determined in the Tone condition of experiments 1 and 2 (Figure 34 and Figure 33, respectively) due to reinforcing, congruent input from the response and distractor in both the TT and VOT fields. The effect of these compatible inputs are visible in both trials as the increase in slope of the activation rise of the Voicing and TT fields in both trials upon the introduction of the distractor input. Both fields in both experiments therefore reach criterion much earlier in the Identity condition than in the Tone condition.



**Figure 38. Comparison of the activation level evolutions in the Identity case. The two depictions show the evolution of a hypothetical Identity case for otherwise unchanged designs of experiment 1 (A) and experiment 2 (B).**

One aspect of the evolutions in the Identity condition to note in both experiments is that it is the Voicing field that reaches criterion second in both experiments. In experiment 1, this is partially due to the pre-shape differences between the Voicing and TT fields. In both experiments though, the rate of rise of the Voicing field is slightly slower than the rate of rise of the TT field. This is due to the different

weights assigned to the distractor articulator ( $d_{artic}$ ) and voicing ( $d_{voice}$ ) inputs in the model, the reasons for which are addressed in detail in section 6.6.1. This difference does not result in any qualitative difference in RTs in this condition.

A few words are in order here to address in what way it is appropriate to describe the shorter RTs in the Identity condition as resulting from “excitation”. In DFT in general and in the present model in particular, excitation has a specific technical meaning, which is the increase in the activation level of a range of parameter values due to the action of the interaction term (as defined in Chapter 5, section 5.3.2.5). This excitation starts as activation values approach the within-field interaction term threshold (in this model,  $\theta = 0.7$ ), but increases when they exceed  $\theta$ , visible in the slight increase in the rate of rise in the TT and Voicing fields in Figure 33A and Figure 34 when the activation level crosses  $\theta$ . The same excitation is present in the example trials shown in Figure 38, but it is not visually discernible because the point at which the within-field threshold is crossed coincides roughly with the time at which the distractor input is introduced to the field, at time step 200. The increased rate of rise of the activation fields is largely due to the reinforcing effect of the compatible inputs. It would be more accurate in the terms of DFT to describe an effect of two identical inputs as “reinforcement” rather than “excitation”, since the rate of rise would increase with two identical inputs even without any influence of the interaction term. However, the excitation in the strict DFT sense is also important in these trials due to the nature of input in the DFT model. The means of two input

distributions do not need to be the same in order to reinforce each other. Overlapping distributions thus have a reinforcing effect, but this local excitation of the DFT interaction term has the effect of making local distances (as determined by  $\sigma$ ) between input values largely irrelevant, since they excite each other.

#### 6.4. Simulations

This section presents the results of model simulations of the response-distractor task. For each experiment simulation, the model ran 150 simulated trials for each of four conditions (Identity, Tone, Congruent, and Incongruent) at each of 3 SOAs (90, 100, and 110 time steps) for a total of 1800 trials per experiment. On each trial, the time step at which the Monitor determined the RT was recorded, as were the production parameter values chosen on the trial. The activation level of each planning field was reset to its trial-initial state at the beginning of each trial.<sup>4</sup>

On all trials, the response was *ta*. Therefore, the distractor was always *ta* in the Identity condition, *pa* in the Congruent condition for experiment 1, *da* in the Congruent condition for experiment 2, *ba* in the Incongruent condition for experiment 1, and *ga* in the Incongruent condition for experiment 2. There were more actual responses and distractors in the experiments than were included in the simulations. However, the behavior of the model was the same regarding cross-field inhibition

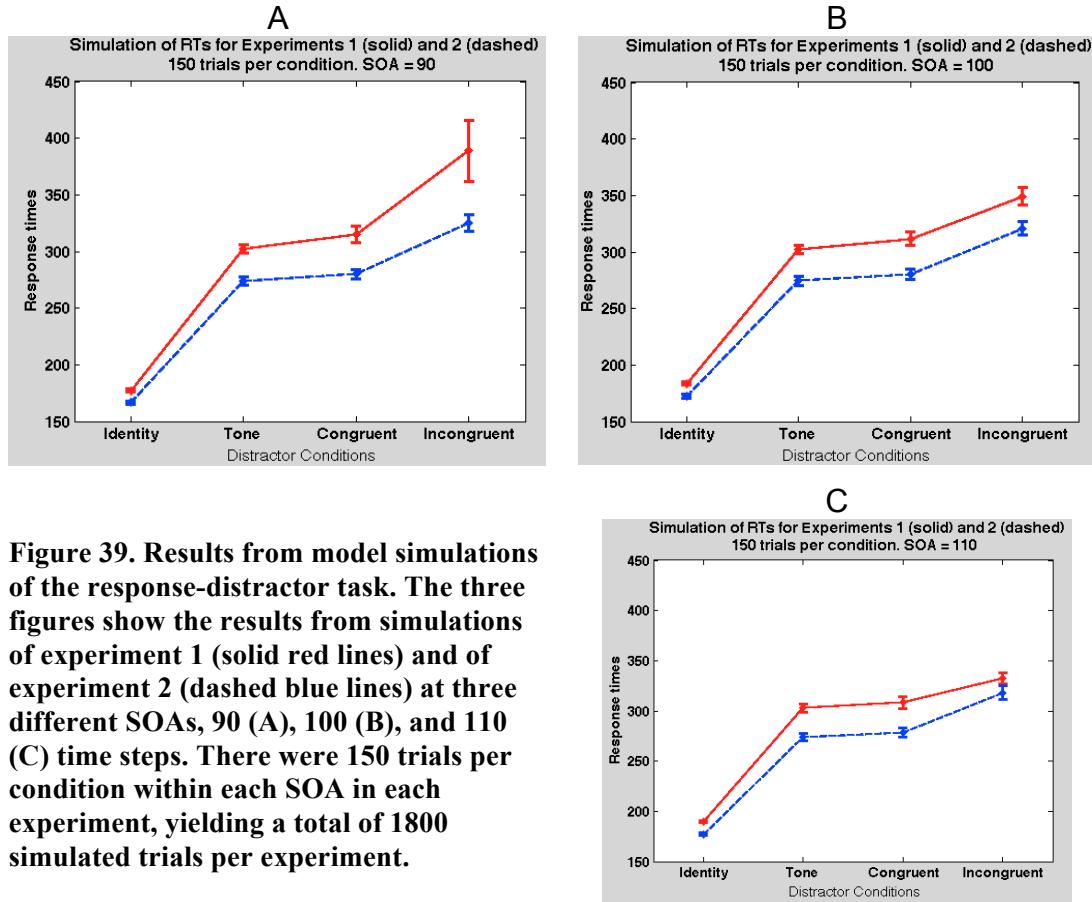
---

<sup>4</sup> The trial-initial states for the fields were stipulated to start at time step 1 at roughly the activation level that they would eventually stabilize at without any response or distractor input, i.e., at the resting level  $h$  + noise + pre-shape. It was implemented this way for display purposes and had no practical effect on the evolution of the field, since the activation levels would soon stabilize at these levels.

irrespective of which particular sets of articulator planning fields were involved because the cross-field inhibition was fully symmetrical, the dynamics and variable values governing the evolution of all three articulator fields were identical, and the input values to each field were defined the same way. Similarly, the effects of the within-field dynamics between VOT values in the Voicing field were fully symmetrical, that is, voiced VOT values had the same effect on voiceless VOT values as vice versa. There was therefore nothing to be gained by varying the simulation to reflect the articulator of the response in the simulations: the results would not change in any meaningful way. The script that ran the experiment simulation can be found in Appendix D.

The results of the simulation are shown in Figure 39. The model produced the correct output (VOT and articulator) on all trials. The results of the model simulations replicated the RT results from the experiments reported in Chapter 3. Most important for the account of the present experimental results was that in the simulations of both experiments, RTs in the Incongruent condition were longer than RTs in the Congruent condition at all three SOAs, and RTs in the Congruent case were longer than the Tone condition. In addition to capturing the difference in RTs between the Congruent and Incongruent case, the model also replicated the finding of Galantucci et al. (2009) that RTs in the Identity condition were shorter than in the Tone condition. The effects of SOA and presence of a distractor were attributed to some other set of processes associated with the task that were assumed to be independent of the effects accounted

for in the model (see Chapter 4, section 4.2.1). Including three different SOAs in the simulation was therefore not intended to replicate the effect of RTs increasing with SOA, but rather to show that the effects of Congruency did not crucially depend on a specific SOA value.



**Figure 39. Results from model simulations of the response-distractor task. The three figures show the results from simulations of experiment 1 (solid red lines) and of experiment 2 (dashed blue lines) at three different SOAs, 90 (A), 100 (B), and 110 (C) time steps. There were 150 trials per condition within each SOA in each experiment, yielding a total of 1800 simulated trials per experiment.**

The model simulations also replicated an unusual aspect of the experimental results in Chapter 3, namely, that the RTs in experiment 1 were shorter than those in experiment 2, as can be seen by comparing Figure 5 in Chapter 3 with Figure 10 in Chapter 3. The mean of the actual RTs across all SOAs and conditions in experiment 1

was 548 ms, but in experiment 2 it was 496 ms. This was a difference of 52 ms, much larger than any of the other effects reported in Chapter 3. This held true in both the experiments and the simulations by and large regardless of condition and regardless of SOA, though condition and SOA did interact in that the difference in RTs between the two simulated experiments decreased as SOA increased in the Incongruent condition. The simulated RT differences arose from the difference in the activation level pre-shapes of the Voicing field in the two experiment simulations, which can be seen by comparing the activation levels of the Voicing field in Figure 31A and C. As noted in sections 6.3.1 and 6.3.4, the maximum activation level of the pre-shape of the Voicing field in experiment 1, where the voicing of the response was not known, was lower than the maximum activation level of the pre-shape in the simulations of experiment 2, where the voicing of the response was always known. The lower maximum activation value in the simulations of experiment 1 was due to within-field inhibition introduced by incompatible pre-shape inputs to the Voicing field. Recall that the RT on a given trial depended on the activation level of some  $x$  value of both the Voicing field and one articulator field reaching criterion ( $\kappa$ ), whose value was always the same. Independent of any effects of congruency, the Voicing field simply had further to rise to reach criterion in simulated trials of experiment 1 than in simulated trials of experiment 2. The longer RTs in the simulation of experiment 1 were attributable to the extra time it took the Voicing field to cover the distance between its pre-shape level compared to the pre-shape of the Voicing field in experiment 2.

## 6.5. Model variable values

This section details all of the variables of the model that were used in the examples in section 6.3 and in the simulations reported in section 6.4. The variable values used for the evolution of the DFT fields were:  $\tau = 150$  and  $h = -3.25$ . The noise term was included in each step of the evolution to introduce stochastic behavior of the model across trials. This noise term added/subtracted a random amount of activation averaging approximately 1.25 activation units to every  $x$  value. The resting level of an activation field was therefore about  $-2$  activation units, equal to the resting level  $h$  plus noise (see, e.g., the LL field level in Figure 31A and C). The values for the interaction term were the same in all four activation fields:  $\theta$  (interaction term threshold) = 0.7,  $w_{excite} = 0.45$ ,  $w_{inhibit} = 0.1$ ,  $= 1$ ,  $\sigma = 1$ . For the sigmoid function,  $\beta$  was always 1.5.

All inputs—pre-shapes, responses, and distractors—to the activation fields were normal distributions of activation across a range of parameter values, as defined in (7) in section 6.2.2. The constriction location input distribution for all articulator fields had a mean of 0 and SD = 2, defined on an arbitrary scale of constriction locations that ranged from  $-10$  to  $10$ . For the Voicing parameter,<sup>5</sup> distributions for all voiced stimuli input had a mean of 5 ms VOT and SD = 45 ms. Distributions for all voiceless stimuli input had a mean of 55 ms VOT and SD = 45 ms.

---

<sup>5</sup> The VOT scale was also implemented in the MATLAB script on a scale of  $x = -10$  to  $10$ . Millisecond equivalent values of VOT were calculated by the following formula:  $x_{ms} = ((x + 10) * 5) - 20$ . This represents a smaller window of VOT values than were used in Chapter 5, but this has no material effect on the model.

The amount by which a given input influences a planning field in the model depends on two properties of that input, all other things held equal. The first is the strength of the input, that is, on a given time step in the evolution of the field, how much the input adds to the activation level of a range of  $x$ -values. The second is the duration of the input, that is, for how many time steps that input is added to the field in its evolution. These two properties are independent of each other and are manipulated differently across the three inputs to the model.

The pre-shape weight ( $p$ ) was 0.8, which applied equally to the inputs to the articulator and Voicing fields. As discussed in section 6.3.1 and shown in Figure 31, the pre-shapes persist as peaks of a low level of activation throughout the evolution of the fields in the absence of any other inputs. The weight of the pre-shape input is very small relative to the other two input types. This low level of activation is deliberately chosen to be insufficient for the fields to stabilize above the resting level via the DFT dynamics. The constant pre-shape level is achieved by having the pre-shape input persist throughout the entire evolution of the field. That is, the duration of the pre-shape input is always the entire evolution of a field. Without this constant input, the activation levels of all fields would drop to the resting level of the field (around -2), and would be no different from the fields without any pre-shape activation, e.g., the LL field in the examples from both experiments shown in Figure 31.

The response input weight ( $r$ ) was 2.7, and was the same for inputs to both the articulator and Voicing field of the required response. As indicated in section 6.2.2, a

small amount of noise was introduced to the actual input mean on each trial for articulator constriction location and VOT for the responses. All response inputs started on time step 100. This is largely for the benefit of displaying the evolution of the fields in the figures in section 6.3. Having 100 time steps made the pre-shapes of the fields visible in the comparison figures (e.g., Figure 33A). The response duration was 400 time steps.

There were two different distractor input weights, one for the articulator parameter ( $d_{artic}$ ), which was 7.5, and one for the voicing parameter ( $d_{voice}$ ), which was 6.3. This difference is due to the fact that the dynamics that give rise to the within-field and cross-field inhibition are markedly different, and is addressed in detail in section 6.6.1. All distractor inputs started on the time step equal to 100 (the start of the response input) plus the SOA value on that trial. The input to the articulator and Voicing fields had the same duration of 125 time steps.

The cross-field inhibition threshold  $\chi$  was 0. The amount of cross-field inhibition ( $q$ ) subtracted on each step from other fields when an articulator field was above was 1.25. The criterion level ( $\kappa$ ) at which the Monitor chose the production values to send to implementation was 6.

A discussion of the values of these variables follows in section 6.6.1. The script that ran the individual trial simulations is found in Appendix C.

## **6.6. Discussion**

The main result of the model is that it provides an account of the differences found in the experiments reported in Chapter 3. The model provides a formalization of a specific link between perception and production, namely, that phonological parameters of a perceived utterance obligatorily serve as input to the planning process of an utterance to be produced. The results from Chapter 3 show that effects of congruency are not limited to the Identity condition as found by Kerzel and Bekkering (2000) and Galantucci et al. (2009), as RTs were shown to be sensitive independently to the properties of voicing and articulator. In light of these experimental results, the Identity condition seems qualitatively different from the Congruent and Incongruent conditions of Chapter 3 in that only RTs in the Identity condition were shorter than in the Tone condition. However, no special computational mechanisms are needed in the DFT model presented here to account for this result for the Identity condition.

The rest of this section first addresses and motivates the specific variable values used in the model. A discussion follows as to what degree the implementation of the model here is assumed to be general to normal speech production vs. specific to the experimental task. The specific results of model are then discussed, including the effects of (in)congruency reported in Chapter 3 and the differences in RTs between the experiments in Chapter 3. The last sub-section presents new, experimentally testable predictions that are made by the model.

### **6.6.1. Determining variable values**

The precise variable values of the model are not particularly meaningful on their own. The values are best interpreted relative to all of the other variable values in the context of the dynamics of the model. In this regard, the strength and duration of the response input can serve as a reference point for interpreting the other values in the model. The duration of the response input may seem unnecessarily long since the Monitor determined the parameter values before the end of the response input in all of the example trials illustrated in section 6.3. This choice of duration reflects the assumption that the response input can and does persist in the planning for the entire duration of a trial because the subject can hold the required response in memory for the entire trial. This assumption seems reasonable, not least because the visual cue indicating the response remained on screen until a response was detected. The response input duration is kept long so that no modulations in RT would potentially depend on whether the response input was still present. It also is not clear how to make a principled decision as to when the response input should cease.

Distractor inputs are distinct from the response inputs in three ways: the distractor inputs are shorter and stronger, and differed in the strength between their Voicing and articulator input. These differences are driven by the experimental results. Specifically, while subjects' RTs showed a significant sensitivity to the various experimental conditions, they did by and large always produce the response that was required on a given trial. In experiment 1, 1382/20481 (7%) responses were errors,

and in experiment 2, 1949/35266 (6%) responses were errors.<sup>6</sup> These error numbers include trials on which the subject replied with the other possible response of the block (though see section 3.2.2 in Chapter 3 on the inclusion of these trials in the analysis), which were the most common, as well as yawns, nonsensical responses, “corrected” responses (e.g., saying “paata” when the answer should have been *ta*), and saying nothing. It was extremely rare for a subject to produce the distractor rather than the required response: just 16 total trials in experiment 1 and 19 total trials in experiment 2. Therefore, a plausible model has to reliably choose the correct response on a given trial in addition to accounting for the RT modulations. Modeling distractors as relatively short but forceful disruptions to the planning process achieves that goal.

Figure 37 illustrates these requirements. In order for incompatible distractor inputs to slow down RTs, the distractor inputs either have to get sufficiently close to  $\theta$  to increase within-field inhibition or reach  $\chi$  to introduce cross-field inhibition, depending on whether it was the Voicing or articulator that mismatch. On the other hand, they cannot be so strong or last so long that they override the response input and resulted in erroneous output. The activation levels of the TT and TB articulator fields in Figure 37 show that the distractor TB input has to overcome two hurdles to reach  $\chi$ . First, even when the two fields start from the same trial-initial level (Figure 37B), the distractor input to the articulator field has to overcome the “head start” that the

---

<sup>6</sup> Total trials do not include those where the SOA error introduced by the presentation software was greater than 30 ms.

response input has since the distractor input always starts after the response input. This handicap is worse on trials in blocks where the distractor articulator is not a possible response and therefore has a lower trial-initial activation level than the response articulator (Figure 37A). Second, because the TT field level has time to rise to  $\chi$  from the response input, the TT field has already begun to inhibit the TB field. The distractor articulator strength was chosen so that it would be strong enough to overcome these handicaps, yet not so strong that it would inhibit the response articulator field so much that the wrong articulator would be chosen for production by the Monitor.

Similar considerations dictate the choice of the distractor Voicing input strength value ( $d_{voice}$ ). The amount of activation introduced to the Voicing field by the distractor (Figure 36) has to be sufficiently strong to slow down the evolution of the peak of VOT activation for the response, but not so strong as to have the distractor VOT value activation level overpower the response VOT value activation, which would result in the wrong VOT being selected by the Monitor. The inhibition introduced by incompatible VOTs is different from the inhibition introduced by incompatible articulators in two ways. First, within-field inhibition arises gradually as some activation value(s) approach the soft threshold  $\theta$  of the interaction term. This is in contrast to the cross-field inhibition, which is triggered by a hard threshold  $\chi$ . This means that if  $\theta = \chi$ , the effects of within-field inhibition are seen sooner than those of cross-field inhibition. This difference explains why within-field inhibition can be seen

at lower levels of field activation than when cross-field inhibition is seen, even though  $\theta > x$ . Second, since every value of  $x$  whose activation level is sufficiently close to  $\theta$  inhibits the rest of the field by  $w_{inhibit}$ , there is a cumulative effect of inhibition as peaks of activation rise toward  $\theta$ . This is in contrast with cross-field inhibition, which is triggered by a hard threshold that subtracts a constant amount ( $q$ ) from the other fields once per time step, if any value of an articulator field is above  $\chi$ . In light of these differences in dynamics, the amount of activation that is subtracted from the VOT field at each time step by each  $x$  value ( $w_{inhibit} = 0.1$ ) is much smaller numerically than the amount of cross-field inhibition (1.25) to achieve roughly comparable inhibition effects within and across fields. The differences in dynamics and inhibition values also requires a different and lower weighting for the distractor voicing input than for the distractor articulator input. Note that this difference in weights is not meant to imply that the articulator of the distractor is more perceptually salient than the voicing of the distractor. The model does not address how the perceived acoustics were transduced into the parameter values involved in the model. The difference in distractor input weights ( $d_{artic}$  vs.  $d_{voice}$ ) reflects only the relative influence of those inputs within the model in the context of two different inhibitory dynamics.

The level of the cross-field inhibition threshold ( $\chi$ ) and the strength of the cross-field inhibition are also driven by the experimental results. Figure 35 illustrates that cross-field inhibition has to be sufficiently strong to slow down the rate of rise of the articulator field, but not so strong as to suppress the activation level of the

articulator field for the required response too much, or the articulator field of the distractor would potentially reach criterion before the articulator field of the required response. This is undesirable because subjects virtually never replied with the incongruent distractor articulator 16 total trials in experiment 1 (in experiment 2 the distractor always matched the articulator of one of the responses, so such errors are impossible to calculate for this experiment).

The choice of the specific criterion value used by the Monitor ( $\kappa = 6$ ) is principled but not crucial. Figure 33A shows that the activation levels of the TT and Voicing fields continue to increase even after criterion has been reached by both fields. This is because the inputs for the required response persisted in the evolution until time step 500. The evolution of the fields is independent of the Monitor function, with the latter having no influence on the former. Figure 33A shows that after the response inputs stops, the maximum activation levels of both fields drops a little, but then stabilizes just above an activation level of 6. This stabilization is due to the dynamics introduced by the interaction term. A  $\kappa$  of 6 indicates that the fields have evolved to a level that, barring any newly-introduced inhibition, the Monitor can be certain the parameter values of the two fields chosen will be representative of the values at which the fields will ultimately stabilize, modulo noise in the evolution, which never qualitatively changes output values.

### **6.6.2. Generality of the model**

It was noted in section 6.1 that the model presented here is assumed to be the normal mechanism by which phonological parameters are set, but that there were also some adjustments in the model that are (or may have been) task-specific. The aspects of the model that are not assumed to be part of the normal process of phonological parameter setting include the variable values of the cross-field inhibition and of the Monitor. In the experimental task modeled here, all of the responses involved syllable-initial stops that have only one primary oral articulator. There are no consonants that required multiple oral constrictions, like, e.g., /w, l, or r/ (in English) or /k<sup>p</sup>, g<sup>b</sup>/ (in, e.g., Yoruba: Ladefoged & Maddieson, 1996). The cross-field inhibition for stops with one primary oral articulator may not be the same as for stops involving multiple oral articulators. In addition, experiment 2 was designed such that when subjects realized they had to produce a stop with one articulator, it was also clear that the other articulators would not be needed. This may have led to an increased level of cross-field inhibition than in normal speech production, though this was not tested. The statement that the cross-field inhibition values used in the present model are not assumed to be the same as in the general case of normal speech production simply is meant to reflect that how and whether their values change in different contexts remains an open question.

As far as the Monitor, the value of the criterion variable ( $\kappa$ ) may well be task-influenced if not task-specific. In both experiments reported in Chapter 3 and the

experiment of Galantucci et al. (2009), subjects were instructed to reply as quickly as they could after the display of the cue indicating the response on that trial. It seems reasonable to expect that some subjects at least may have lowered the criterion value to be able to reply as soon as they were reasonably confident of the response on a given trial. This level may have been lower than what it would be in normal speech production. It is also likely that in a different task, the Monitor might require a different way of choosing values to send to production. The Monitor could, for example, be forced to choose production values at a particular point in time rather than based on activation levels. Yuen, Brysbaert, Davis, and Rastle (2010) present a task where this may plausibly have been the case. In their task, subjects had to produce disyllabic nonsense response utterances (e.g., “a seeb”, “a keeb”) based on a visual cue, which was presented immediately following an auditory distractor. Subjects had to respond in sync with a beep that followed 500 ms after the presentation of the cue. Data were collected using electropalatography. The results of interest were that *s*-responses (“a seeb”) preceded by *t*-initial distractors (“teeb”) showed increased alveolar contact compared to the same responses with *s*-initial distractors (“seeb”). In addition, *k*-responses (“a keeb”) preceded by *t*-initial distractors (“teeb”) also showed increased alveolar contact compared to the same responses with *k*-initial distractors (“keeb”). Assuming the results of Yuen et al. (2010) were the result of increased activation levels like those in the present model, which is their interpretation of their results, they would be compatible with a Monitor function that took all activation

levels at a set point in time and sent them to Implementation, rather than waiting for the levels to reach a particular activation criterion or stabilize. Looking at Figure 37A for example, if instead of functioning as in the present model the Monitor took all positive activation levels at, say, time step 400, then constrictions of two articulators (TT and TB) would be specified for implementation. This is similar to what Yuen et al. (2010) found in the “keeb-teeb” case. If the parameter axis shown for the articulator fields in the present model were constriction degree rather than constriction location (cf., Figure 30), then the influence of a distractor (“teeb”) with a greater value for constriction degree of the same articulator (TT) as the response being planned (“seeb”) could be accounted for by the influence of the distractor input on the evolving TT field. Therefore, the function of the Monitor is assumed to be task-specific. Though the values of both the cross-field inhibition and the Monitor are likely task-specific, the mechanisms themselves are assumed to be part of the normal process of planning since it is beneficial to suppress the activation of unwanted articulators in general, and values must at some point be chosen for Implementation, whatever the criteria.

### **6.6.3. Incongruity effects**

The main result of the model is that it provides an account of the difference in RTs between the Congruent and Incongruent conditions of experiments 1 and 2 from Chapter 3. The differences in RTs simulated by the model arise from two different

sources in the model, within-field and cross-field inhibition. While the results of the model in the various conditions reported in section 6.3 show how the two types of inhibition can explain the RT modulations observed in the experiments, a puzzle presents itself in the Incongruent cases. Figure 40 summarizes the inhibition that is introduced in each condition in the simulations of each of the two experiments.

|    | <b>Exp.</b> | <b>Condition</b>   | <b>Mismatch</b>     | <b>Inhibition</b>                    |
|----|-------------|--------------------|---------------------|--------------------------------------|
| A) | 1           | Congruent          | articulator         | cross-field                          |
| B) | 1           | <i>Incongruent</i> | articulator + voice | cross-field<br>within-field          |
| C) | 1           | Incongruent        | articulator + voice | cross-field<br><b>X</b> within-field |
| D) | 2           | Congruent          | voice               | within-field                         |
| E) | 2           | Incongruent        | voice + articulator | within-field<br>cross-field          |

**Figure 40. Sources of inhibition in the Congruent and Incongruent conditions in the two experiments, assuming that within-field inhibition imposes a slightly smaller slow-down on RTs than cross-field inhibition, as indicated by the width of the boxes.** A) shows that the Congruent condition in experiment 1 was due to cross-field inhibition due to mismatching articulators between the response and distractor. D) shows that the Congruent condition in experiment 2 was due to within-field inhibition due to mismatching voicing. B) show that the slow-down in the Congruent condition cannot be due to the two types of inhibition being additive, since the effects of cross-field inhibition would obscure the effects of within-field inhibition. C) shows that there must be some other contributing source ('X') of slow-downs in RTs in the Incongruent condition of experiment 1.

Within-field inhibition is a crucial factor in the outcome of the model when there is a mismatch in VOT between the two inputs, which is the case in the Congruent condition of experiment 2. Cross-field inhibition is a crucial factor when there is a mismatch in articulator, which is the case the Congruent condition of experiment 1. Both types of inhibition are therefore introduced into the field

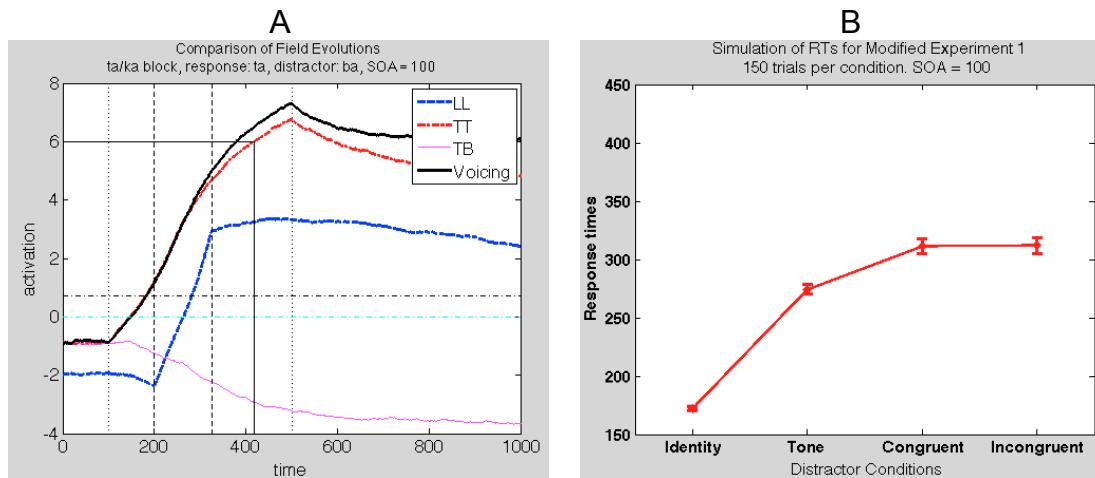
evolutions in the Incongruent case of both experiments, as in both experiments the distractor always differed from the response in articulator and voicing. If the effects of the two types of inhibition were simply additive, then the outcome of the model would be straightforward. The effects of within-field only or cross-field inhibition only in either Congruent condition would be less than within-field only or cross-field in the Incongruent condition.

However, the two inhibitions cannot and do not have any direct effect on each other, since there is no interaction between the Voicing and articulator field evolutions. The effects of inhibition therefore are not additive (see section 6.3.5). How the model yields a difference between the Congruent and Incongruent conditions in both experiments is not immediately obvious. Assume that cross-field inhibition introduces a slow-down of amount X, and within-field inhibition introduces a slow-down of amount Y, where  $X > Y$ , represented in Figure 40 by the width of the boxes indicating the relative amount of the two types of inhibition. The Congruent condition in each experiment (Figure 40A and D) is easily obtained, as the inhibition present in a given experiment should result in RTs being longer than the Tone condition, where there is no inhibition present. The difference in the amount of slow-down introduced by the two types of inhibition also straightforwardly captures the fact that the RTs in the Incongruent condition in experiment 2 were longer than in the Congruent condition, as the slow-down due to the cross-field inhibition is greater than the slow-down due to the within-field inhibition (Figure 40E). The puzzle then is how these two

types of inhibition can account for RTs in the Incongruent condition of experiment 1 being longer than in the Congruent condition. Since the within-field inhibition is less than the cross-field inhibition, the cross-field inhibition is still going to be the decisive factor for determining RTs, as it will always cause the articulator field to reach criterion second. Even though within-field inhibition will be present, its effects should never be seen because they will be obscured by the effects of cross-field inhibition (Figure 40B).

The answer to the puzzle lies in the pre-shape of the fields in experiment 1. Due to the lower trial-initial activation level of the Voicing field in experiment 1 (Figure 37A), the evolution of the Voicing field not only has to overcome the influence of the within-field inhibition to reach criterion, it also has to do so starting from a lower point of activation than the articulator field. The handicap of the lower pre-shape level (indicated in Figure 40C by the box labeled “X”) plus the within-field inhibition introduced by the incompatible voicing of the response and distractor combine to make the Voicing field reach criterion after the articulator field, and therefore making the RTs longer in the Incongruent condition than in the Congruent condition.. This is similar to the effect of the pre-shape on the simulation of the Tone condition in experiment 1 (Figure 34) versus experiment 2 (Figure 33). This means that the source of the pre-shape “handicap” is also within-field inhibition, since that is the cause of the lower activation level of the Voicing field pre-shape in experiment 1 (see section 6.3.1) Figure 41 illustrates that the difference in simulated RTs between

the Incongruent and Congruent conditions in experiment 1 goes away when within-field inhibition is preserved, but the pre-shape of the Voicing field is not lower than the articulator field.



**Figure 41. Simulation of a modified version of experiment 1, where the potential responses are *ta* or *ka*, and response is *ta*, the distractor is *ba*, and the SOA is 100 time steps. (A) shows that in the Incongruent condition, the within-field inhibition of incompatible distractor voicing is not sufficient to cause the Voicing field activation to reach criterion second. (B) shows the simulations of 150 trials of this modified version of experiment 1. There is no difference between in the RTs between Congruent and Incongruent conditions, unlike in Figure 39.**

The simulations in Figure 41 are a modified version of experiment 1, where the potential responses are *ta* or *ka* instead of *ta* or *da*. Figure 41A shows a single simulated trial of the Incongruent condition of this modified experiment 1. Since the two potential responses have compatible VOTs, the pre-shape of the Voicing field is at the same level as the articulator (TT) field reaches criterion second and determines the RT of the trial since the cross-field inhibition is greater than the within-field inhibition (in

contrast with the simulated trial of the unmodified experimental design shown in Figure 37A). Figure 41B illustrates the results from the simulation of 150 trials of this modified design, showing that the difference between the Congruent and Incongruent conditions went away without this difference in pre-shape.

In sum, the difference in RTs between the Incongruent and Congruent conditions in experiment 1 is due to the introduction of within-field inhibition with two different sources. The within-field inhibition introduced by the incompatible response and distractor VOTs plus the within-field inhibition present in the pre-shape of the Voicing field combine to make RTs in the Incongruent condition of experiment 1 longer than the Congruent condition.

#### **6.6.4. Unknown voicing vs. unknown articulator**

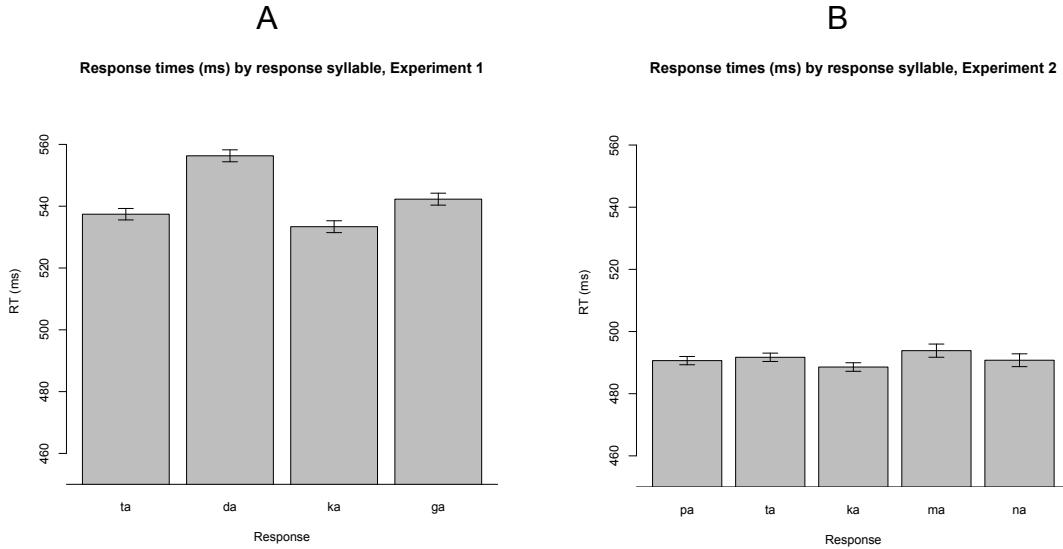
A noteworthy result of the model is that it accounts for the difference in actual RTs across the experiments 1 and 2 from Chapter 3, as noted in section 6.4. Subjects took longer to reply when they did not know the voicing of a reply but knew the articulator (e.g., *ta-da*), than when they did not know the articulator but knew the voicing (e.g., *ta-ka*). A comparable result has not been reported in the literature. The closest result comes from Whalen (1990), which reported results from a naming task where subjects saw an orthographic prompt indicating a nonsense  $V_1CV_2$  utterance where the C or  $V_2$  was not known (e.g., “A\_I” or “AB\_”). When the computer detected the start of  $V_1$ , it would then indicate the missing segment (e.g., “A\_I”

became “API” or “ABI”; “AB<sub>\_</sub>” became “ABI” or “ABU”). RTs were longer when C was not known than when V<sub>2</sub> was not known. C was always either /p/ or /b/. The Whalen (1990) results are similar to the present results insomuch as the articulator was known but the voicing was not. However, there was no condition where the subject knew what the voicing would be but not the articulator (e.g., “A\_I” becoming “API” or “ATI”), so that result cannot establish that it was the unknown voicing of the C that caused the longer RT. It seems more reasonable to assume that the delay reflected the fact that an unknown C was earlier in the utterance than an unknown V<sub>2</sub>. The difference in RTs across the present experiments was in itself a novel finding.

It does not seem to be the case that these differences are instrument or measurement artifacts. The data for both experiments were collected using the same instruments in the same lab, and data labeling was performed by the same experimenter (the author) using the same methods outlined in section 3.2.2 of Chapter 3. Given the large number of subjects in each experiment (38 in experiment 1 and 35 in experiment 2) and given that all subjects were drawn from the NYU Psychology Subject Pool, it seems unlikely as well that the differences are attributable to some characteristic of one subject group compared to the other. Anecdotally evidence has suggested that sometimes experiments run earlier in a semester tend to have faster response times (Carol Fowler, personal communication) compared to those that are run later in the semester, presumably because the former group are more motivated

than the latter. Even if this were true, it could not account for the differences reported here since experiment 1 was run before experiment 2 in the same semester.

It would be wrong to describe this difference in RTs across the present experiments as a novel finding, or to attribute these cross-experiment RT differences to the pre-shapes of the planning fields in the present model, if it simply reflected inherent differences in the production times for the onset consonants of the responses. Rastle, Croot, Harrington, and Coltheart (2005) tested for such inherent differences using a delayed naming task in which subjects produced syllables with various onset consonants. Subjects were shown an orthographic prompt for a monosyllabic utterance. Subjects then pressed a button when they were ready to produce the utterance. The button press triggered a tone, which was the subject's cue to produce the utterance. RTs for stop-initial responses were measured as the time between the tone and the release of the oral closure of the stop, comparable to the measurement used in the experiments reported in Chapter 3 (see section 3.2.2). Rastle et al. (2005) found that stop-initial CV syllables had longer RTs when the stop was voiced than when it was voiceless (replicating a result reported by Fowler, 1979), and that nasal-initial syllables had shorter RTs than stop-initial syllables. This is a potential problem for the experiments reported here in Chapter 3, because the responses in experiment 1 had voiced and voiceless responses, while the responses in experiment 2 had voiceless and nasal responses. The longer RTs in the present experiment 1 could therefore have been due to the inclusion of voiced stops and no nasals.



**Figure 42. RTs by response syllable. The RTs across all SOAs and conditions for experiment 1 (A) and experiment 2 (B).**

Figure 42 shows the RTs from experiments 1 and 2 across all SOAs and conditions. On the one hand, the results from the present experiments corroborated the findings of Fowler (1979) and Rastle et al. (2005). Figure 42A shows that in experiment 1, voiced responses (*da, ga*) had longer RTs than their voiceless equivalents (*ta, ka*).<sup>7</sup> However, both experiments included *ta* and *ka* responses, and they were markedly longer in experiment 1 than in experiment 2, on the order of about 45 ms (compare Figure 42A and B). Therefore, inherent differences in RTs based on properties of the initial consonant of the response syllable could not have been the primary source of the difference in RTs between the two experiments.

<sup>7</sup> This difference between voiced and voiceless responses does not pose a problem for the results concerning the effects of distractor-response congruency reported in Chapter 3. This was controlled for in the experimental design by ensuring that every subject had an equal number of voiced and voiceless responses in each block of experiment 1. It was also controlled for in the statistical analysis by including the response (Item) as a random factor in the statistical model.

It is fair to claim then both that this result from the present experiments is a novel result, and that the model provides a principled reason for this difference between the RTs in the two experiments. Note that no special variable values are introduced to the pre-shapes to get the difference across experiments. The values given in section 6.5 and discussed in section 6.6.1 are driven by the observed effects of congruency in the experimental data. The pre-shapes are included because it is common practice in DFT models (e.g., Erlhagen & Schöner, 2002; Kopecz & Schöner, 1995; Schutte, Spencer, & Schöner, 2003; Thelen et al., 2001) to include task-effects such as these, as they have been shown to have effects on the behavior of the model.

### **6.6.5. Additional predictions**

The model as it stands makes additional specific predictions that could be tested using the same experimental task presented here. One prediction involves a case based on the Galantucci et al. (2009) study. Their experiment 2 was similar to the present experiment 2, in that the response voicing was always known, though in their experiment the response was always voiced (*ba* or *da*). The present model predicts that if their experiment were run where the response-distractor pairs instead always matched in articulator but not in voicing (e.g., *ba-pa* or *da-ta*), the difference between the mismatch and identity conditions should still be found. This result is expected regardless of whether the identity condition is present in the experimental design.

Another prediction of the model is that there should be no difference between the Congruent and Incongruent conditions in an experimental block like the one described in section 6.6.3. This prediction can be tested by having the potential responses in a block match in voicing but differ in articulator (e.g., *ta* or *ka*, as in experiment 2). Distractors would differ in articulator and match in voicing in the Congruent case (e.g., *ta-pa*). Distractors would differ in articulator and in voicing in the Incongruent case (e.g., *ta-ba*). While RTs in the Congruent and Incongruent conditions are not predicted to be different, they together are predicted to be longer than RTs in the Tone condition. See section 6.6.3 for details.

One more prediction of the model is that any within-field manipulations should yield effects comparable to those obtained here for VOT. For example, the difference in the RTs across the experiments from Chapter 3 was due to the fact that the Voicing pre-shape could not rise as high as the articulator pre-shapes because of the within-field inhibition introduced by having two peaks in the same field due to the soft nature of the within-field inhibition threshold (section 6.3.1). This inhibition is not introduced when the pre-shapes are peaks in different fields. The articulator field in the present model is defined as having constriction location as the parameter axis. The model therefore predicts that RTs in an experiment where the two possible responses on a given trial share articulator (e.g., *sa* or *ʃa*) but differ in constriction location should be longer than an experiment where the two responses differ in articulator (e.g., *sa* vs. *ʃa*), even if voicing is known in all conditions. The model predicts this result

because the pre-shapes of *sa* and */a* would be limited in how high they could rise because of within-field inhibition (analogous to the Voicing pre-shape for *ta-da* shown in Figure 31C), while *sa* and */a* pre-shapes would not be limited by within-field inhibition (analogous to the articulator field pre-shapes for *ta-ka* shown in Figure 31A).

## 6.7. Conclusion

This chapter presented a formal model of the timecourse of phonological planning, during which the phonological production parameters for an utterance are set before being sent to implementation. The model formalizes the link between perception and production by having the parameters of a perceived utterance serve automatically and obligatorily as input to the ongoing planning of a utterance to be produced. The model components are assumed to be the general mechanisms by which these parameters are set in speech, though the model as implemented here has certain values that may be specific to the response-distractor task.

The model accounts for experimental results obtained in Chapter 3, where response times were sensitive to the degree of congruency between a planned utterance and a perceived utterance. The effects of mismatching parameters of articulator and voicing on response times in the experimental results are qualitatively the same in that mismatching in either voicing or articulator resulted in longer response times versus a neutral tone distractor, but not in response times as slow as

when the distractor and response mismatched in both voicing and articulator. The model, however, attributes the slow-downs in response times to different sources of inhibition, depending on whether the mismatching parameter is articulator or voicing. The aspects of the model design that give rise to these two different sources of inhibition also enable the model to account for otherwise unexpected differences in response times across the two experiments reported in Chapter 3. Lastly, the model also accounts for experimental results obtained by other researchers, wherein response times were shorter when a perceived distractor was identical to a planned response than when there was a tone distractor. The same mechanisms in the model that account for the present experimental results account for these results without assigning any special status to this “identity” condition. The model also makes a set of specific, new predictions that could be tested using the same experimental task.

## CHAPTER 7: DISCUSSION AND CONCLUSION

### 7.1. Summary

In this dissertation I have presented results from experiments using a response-distractor task that provide evidence for independent perceptuo-motor effects of articulator and voicing (Chapter 3). Response times were shorter when a distractor and response mismatched in either voicing or articulator (e.g., *ba-da* or *da-ta*) than when they mismatched in both voicing and articulator (e.g., *ba-ta*). An analysis of these results in light of other results in the literature shows that any model that proposes to account for these effects must include the computational principles of excitation and inhibition. With these results and analysis in mind, a dynamical model of phonological planning has been presented in Chapter 6. This model formalizes the link between speech perception and speech production as the phonological parameters of a perceived utterance obligatorily serving as input to the ongoing phonological planning of an intended utterance. The model provides a principled account accounts for the congruency effects found in the experiments reported in Chapter 3. It also accounts for perceptuo-motor effects of identity (e.g., *ba-ba*) found by other researchers, and for an unexpected difference in response times across the experiments in Chapter 3 (see section 6.6.4 in Chapter 6). In addition, the computational framework employed in the model was also shown to be able to account for gradient and categorical data from a common phoneme classification task (Chapter 5).

## **7.2. Theoretical implications**

The empirical findings from this dissertation have implications for theories of speech perception, speech production, and how the two are linked. The present results also inform the issue of the status of segmental identity in the psycholinguistic and phonological literature.

### **7.2.1. Perception-production link in speech**

As discussed in section 2.4.2 of Chapter 2, there is considerable debate in the speech perception literature about the role that the speech production system plays in speech perception (Diehl, Lotto, & Holt, 2004; Fowler, 1996; Galantucci, Fowler, & Turvey, 2006; Lotto, Hickok, & Holt, 2009; Ohala, 1996), with the primary point of contention being to what degree speech perception crucially relies on motor/production representations. The effects outlined in sections 2.3.1 and 2.3.2 as well as those found in the experiments reported in Chapter 3 are referred to here and (Galantucci, Fowler, & Goldstein, 2009; Kerzel & Bekkering, 2000) elsewhere as “perceptuo-motor effects” because they are assumed to be the result of production (or motor) codes that are activated during speech perception. In the model detailed in Chapter 6, those motor codes are the phonological parameter values of the perceived utterance, which serve obligatorily as input to any ongoing phonological planning. The claim in this dissertation is not, however, that the codes activated by perceiving the distractor must exclusively be these motor codes. Rather, the claim is that the codes

activated in the perception of the distractor must minimally be these motor codes. Auditory-acoustic or featural codes could also be activated in the perception of the distractor as long as their activation excites the associated motor codes linked to these auditory-acoustic or featural codes. The experiments presented here were not designed to address whether non-motor codes are also activated. According to the Motor Theory of speech perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985), there are no intermediary codes such as acoustic-auditory codes between the acoustic signal and the motor codes: the production system is necessarily activated during speech perception because these motor codes are the sole object of perception in the Motor Theory. The results of the experiments and model presented in this dissertation are fully compatible with the Motor Theory. They are also consistent with theories that do or could propose a link between auditory-acoustic or other types of codes, which are activated during the perception of the distractor, and motor codes corresponding to these auditory-acoustic codes (cf. Viviani, 2002: Figure 21.12, p. 436).

The present account is also consistent with Direct Realism (Fowler, 1986), which is similar to the Motor Theory in that the objects of perception in Direct Realism are speech gestures. Direct Realism differs from Motor Theory, however, in that it explicitly does not “propose a role for the speech motor system in speech perception” (Fowler, 1986, p. 1731). That is, the form of the objects of perception is gestural in Direct Realism, but these gestural representations do not invoke the

analysis-by-synthesis (Stevens & Halle, 1967) provided by the speech production system that is proposed in the Motor Theory. The present results are consistent with Direct Realism under the assumption that the motor codes in the model are activated during speech perception via some link with Direct Realism's gestural representations, even though such activation is not a strict requirement of Direct Realism.

Similarly, the present results are also compatible with theories recognizing or emphasizing the role of acoustic-auditory codes (e.g., Ohala, 1996) or abstract featural codes (e.g., Mitterer & Ernestus, 2008), again, as long as a link between these other codes and the motor codes is assumed, or if they are activated in parallel as proposed, e.g., in the neurophysiological model of speech processing of Hickok and Poeppel (2007). The link between perception and production must be specified at the level of setting production parameter values, including articulator and voicing, that need to be activated either directly or via associated auditory, featural, or other codes. The effects of the link between perception and production are seen as the influence of a perceived distractor on the process of setting those parameters, i.e., on a production process, as witnessed by the RT modulations found in the present and previous experiments. In sum, the present results neither provide evidence against any type of theory of speech perception, nor provide an argument for one type of theory of speech perception over another, so long as the theory in question admits the possibility of the activation of motor codes during speech perception, whether directly or indirectly. The present results do present a challenge to any theory that would not admit the activation of

motor codes during speech perception or that does not accommodate a role for representations at the level of phonological features.

### **7.2.2. Theories of speech production**

The results presented here show that the parameters of voicing and articulator have separable roles in the process of speech planning, as witnessed by the perceptuo-motor effects in experiments 1 and 2 from Chapter 3. These results therefore have implications for theories of speech production, which differ on what role if any they assign to speech parameters in planning (as outlined in section 2.2 of Chapter 2). In the WEAVER++ model of word-form encoding in speech production (Levelt, Roelofs, & Meyer, 1999; Roelofs, 1997, 2000) no critical role is assigned to parameters such as voicing or articulator of individual segments in the preparation of articulation (Roelofs, 1999). Instead, the segments of a word to be produced activate stored articulatory plans of full syllables (following Levelt and colleagues: Levelt et al., 1999; Levelt & Wheeldon, 1994). The importance of syllabic structure in speech production is well supported (e.g., Browman & Goldstein, 1988; Krakow, 1999; Shaw, 2010; Smith, 1995, among many others), but the present results show that the process of phonological planning is also sensitive to the parameters of voicing and articulator. On the other hand, the models of Dell and colleagues (Dell, 1986; Dell, Juliano, & Govindjee, 1993; Dell & O'Seaghda, 1992) and Meyer and Gordon (1985) treat the programming of production parameters (formalized as phonological features

by these authors) as a critical component of word production (see Levelt, 1999, for discussion), an approach that is more consistent with the present results. A model of speech production that does not accommodate planning at the level of the parameters of voicing and articulator cannot account for the perceptuo-motor effects found here. Bohland, Bullock, and Guenther (2009) leave room in their GODIVA model for the potential accommodation of phonological parameter setting, should empirical evidence motivate such an expansion. The present results show that such an expansion is warranted.

### **7.2.3. Identity**

The experimental results from Chapter 3 provide the first unambiguous experimental evidence for perceptuo-motor effects arising from two stimuli that are not segmentally identical. The model presented in Chapter 6 accounts for these effects through a combination of excitation and inhibition introduced by a distractor in the process of phonological planning. Notably, the model also derives the results found by others (Galantucci et al., 2009; Kerzel & Bekkering, 2000) for the condition in the same task when the distractor and response are identical, and it does so without the need to treat the identity condition as special. The same computational principles derive the results in all of the experimental conditions reported: Identity, Congruent (i.e., mismatching on one parameter), and Incongruent (i.e., mismatching on two parameters). The answer to the question of whether the identity case is unique in

giving rise to perceptuo-motor effects is straightforwardly “no”. This conclusion does not undermine the notion that identity may have some privileged status in other phonological or cognitive processes (see section 2.4.1 of Chapter 2 for discussion). In fact, the results from the experimental evidence and the present model show that although no special computational mechanisms are required to derive the effect of identity, RTs in the Identity condition are qualitatively different from the other conditions (see section 6.3.6 of Chapter 6). Only in the Identity condition are RTs faster than in the Tone condition, due to the fact that only in the Identity condition are the effects of local excitation exposed. In this sense the Identity condition is special, and this may be of importance to other processes that are not germane to the dynamics of phonological planning.

### **7.3. Future research**

The model developed in Chapter 6 of this dissertation also raises some questions and makes some predictions that could be addressed with further experiments and development of the model.

#### **7.3.1. Testing additional predictions of the model**

As noted in section 6.6.5 of Chapter 6, the model makes several predictions that were not tested in the experiments in Chapter 3. Specifically of interest would be to test the prediction that response times in an experimental block where the two potential responses differ in activation values defined within a single planning field

should be longer than in a block where the two potential responses involve activation of different planning fields. The effect size for the response time difference between experiments 1 and 2 in Chapter 3 was very large (about 52 ms), compared to the effect sizes found for perceptuo-motor effects, which were between 6-12 ms. Testing this prediction should therefore require far fewer subjects to achieve sufficient statistical power. This opens the practical possibility of testing this prediction using articulatory data collected with electromagnetic articulography ("EMA", Hoole, Zierdt, & Geng, 2003; Perkell et al., 1992) instead of acoustic data, despite the fact that data-collection using EMA is more onerous than for the type of data collected in the experiments in Chapter 3. The benefit of using articulatory EMA data would be that it could provide be a more direct test of the onset of articulation, rather than determining that onset by way of the acoustic signal.

### **7.3.2. “Articulator” effects**

Throughout this dissertation, the effects that were found in experiment 2 in Chapter 3 have been described as effects of “articulator” (or, more specifically, “primary oral articulator”, see footnote 1 of Chapter 2), and it may be the case that these RT effects arose due to the fact that in the Congruent condition the response and distractor shared articulator but differed in articulator in the Incongruent condition. Several previous researchers sought—though uniformly failed—to find facilitative effects based on shared articulator (Galantucci et al., 2009; Gordon & Meyer, 1984;

Mitterer & Ernestus, 2008), with the assumption that two stimuli having an articulator in common should be sufficient to generate perceptuo-motor effects. However, according to the model in Chapter 6, the effect should be more precisely described as the effect of two inputs having the same articulator and where the constriction degree and/or constriction location of that articulator are also the same. In fact, the model presented in this dissertation predicts no facilitative perceptuo-motor effects of articulator *per se*. To the contrary, a response and distractor that share articulator but differ in constriction degree and/or constriction location should not introduce local excitation, but rather introduce within-field inhibition and therefore slow down RTs. According to the model, facilitation arises only when the values of one parameter axis of a given field are the same. This prediction can be terminologically distinguished, using the formalisms of Articulatory Phonology (Browman & Goldstein, 1986, et seq.), by referring to *tract-variable* (Browman & Goldstein, 1990) effects as opposed to *articulator* effects (or, more narrowly, primary oral articulator, see footnote 1 of Chapter 6). That is, the present model predicts facilitation when the tract variables of a response and distractor have the same values, rather than when the two simply involve the same tract variable. This prediction was not tested in the experiments in Chapter 3, so the empirical question remains open. Table XIX provides stimuli for a response-distractor task that would test this prediction, and illustrates why the model makes the prediction it does.

In the experimental block shown in Table XIX, the potential responses (*ʃa-fa*) are both voiceless fricatives that differ in articulator (tongue-tip vs. lower lip, respectively). Distractors differ in articulator from the responses, except for the *ʃa-sa* response-distractor combination (the congruent pair marked with a green ✓ in Table XIX), where they both involve a tongue-tip constriction. Although they share articulator, they differ in constriction location (post-alveolar vs. alveolar, respectively). That is, they both make use of the tract variable of Tongue Tip Constriction Location but differ in the values assigned to it. The responses in this block are similar to the responses from a block from experiment 2 (Table XI from Chapter 3), with two differences. The first difference is that the responses are fricatives instead of stops, i.e., they have a constriction degree of critical instead of closed. This does not have any effect on the predictions of the model, which are affected only by whether the parameter values are the same or not, not by what the

**Table XIX. Stimuli to test predictions of the model regarding articulator vs. tract-variable effects. Responses are either post-alveolar (*ʃa*) or labio-dental (*fə*) fricatives. Distractors always match in manner (i.e., they are all fricatives) and voicing (voiceless). Only in the *ʃa-sa* combination do the response and distractor share articulator, though they differ in constriction location.**

| Response  | Distractor |           |             |             |
|-----------|------------|-----------|-------------|-------------|
|           | <i>sa</i>  | <i>ha</i> | <i>tone</i> | <i>none</i> |
| <i>ʃa</i> | ✓          | X         | •           | •           |
| <i>fə</i> | X          | X         | •           | •           |

values are if they are the same. The second difference is that even though the congruent pair share articulator, they differ in constriction location, which is a dimension of the Tongue Tip articulator field defined in the model (see section 6.2.1 in Chapter 6).

In the congruent condition, the mismatch in constriction location should introduce within-field inhibition, resulting in the TT field evolving more slowly than if that incompatible input were not present. RTs should therefore be slower in this condition than the Tone condition. On the other hand, the hypothesis that effects of articulator *per se* should be found should predict that RTs in the Congruent condition should be faster than in the Tone condition.<sup>1</sup>

### 7.3.3. Phonetic modulation

The model using Dynamic Field Theory (DFT) presented in Chapter 6 makes predictions in addition to response time modulations in the response-distractor task. Recall from section 6.2.1 of Chapter 6 that the planning fields in the model are defined as having one axis for each parameter that needs to have values set for production, i.e., VOT for the Voicing field and constriction location for the articulator fields. In the model, the point at which a peak of activation in a field stabilizes indicates not only

---

<sup>1</sup> The prediction for the Incongruent conditions (marked with a red X in Table XIX) is that RTs should be longer than when there is a tone distractor due to cross-field inhibition introduced by the incongruent articulator of the distractor, assuming for /h/ that the articulator field for the glottis has one dimension for constriction degree and cross-field inhibitory links to the other articulator fields as in section 6.2 of Chapter 6. Whether the incongruent and congruent conditions are different from each other does not bear on the difference between the articulator vs. tract-variable hypothesis, as long as they both have longer RTs than the Tone condition.

the time step at which values are sent to production, but also what the specific production value is. The combination of inputs to the planning process can also result in modulations of the production values of the utterance being planned.

As was illustrated in Chapters 5 and 6, the dynamics of DFT are such that when two inputs are far away within the same field, they inhibit each other. On the other hand, two inputs that are sufficiently close to each other—even if they are not the same—mutually excite each other. This mutual excitation results in faster a build-up of activation for parameter values in the region of the two inputs than if there were no re-enforcing input, thus the increased rate of build up in the good-exemplar cases in Chapter 5 (Figure 10 in Chapter 5) or the effect of congruent voicing in experiment 1 from Chapter 6 compared to hearing a tone (compare the rate of rise of the voicing field in Figure 7 with Figure 6 in Chapter 6). That is, local excitation introduced by sufficiently activated parameter values increase not just their own activation level but the activation level of neighboring parameter values (as defined by the interaction term, see section 5.3.2.5 of Chapter 5). Therefore, given two inputs that are sufficiently close to each other, one having peak a maximum activation at parameter value  $x$  and the other having a maximum of  $x + y$ , all parameter values between  $x$  and  $x + y$  are excited by both inputs. Assuming the inputs are of equal strength and that the combined inputs are of sufficient strength for the field to stabilize with a single peak of activation, the parameter value with the maximum activation level when the field

stabilizes will be a value between  $x$  and  $y$ , specifically,  $x + (y / 2)$ , abstracting away from the influence of noise in the system.

On the other hand, when there are two incompatible inputs to the same field, they do not mutually excite any parameter values that lie between them, they mutually inhibit each other. This means that in the case of two compatible (i.e., close) inputs, the field reaches a stable state with a peak faster than when there is no reinforcing input, but the actual parameter value chosen for output will be an intermediary value between the maxima of the inputs. It also means that in the case of two incompatible (i.e., distant) inputs, the field stabilizes more slowly than when there is only one input, but there is no influence of one input on the other in terms of the parameter value that gets sent to production. This behavior is qualitatively the same as seen in the model of saccade planning developed by Kopecz and Schöner (1995).

The model in Chapter 6 therefore predicts that on trials where the response and distractor are both similar on a given parameter (say, both voiceless) but where the distractor has a non-prototypical parameter value, responses productions should show the influence of that non-prototypical distractor compared to trials where the distractor is incompatible, where no modulation of that parameter is predicted. Additionally, response times in the former case should be shorter than in the latter. To use a concrete example of voicing, assume a speaker normally produces 40 ms of VOT when uttering *ta*, and on average takes 400 ms to respond in the response distractor task when she hears a tone distractor. Across all trials where this speaker produces *ta* but hears a *da*

distractor having 0 ms VOT, her response times on average should be greater than 400 ms, but the VOT of the *ta*'s should still be 40 ms on average. However, across trials with a *ta* distractor having a VOT of 60 ms, her response times should be shorter than 400 ms, but the VOTs for these *ta*'s should be longer than 40 ms, but not 60 ms. Similar predictions would hold for manipulation of constriction location within an articulator field.

In principle, this prediction could be tested with the data collected in experiment 1 from Chapter 3. VOT values for syllable-initial voiceless stops in English differ based on place of articulation, with *p* having shorter VOT than *t*, which in turn has shorter VOT than *k* (Lisker & Abramson, 1964). Therefore, in that experiment where speakers produced *ta* or *ka*, trials with a congruent *pa* distractor, which had a VOT of 40 ms (shorter than the 58 ms average reported by Lisker & Abramson, 1964), the model predicts shortened VOTs for *ta* and *ka* compared to trials with the incongruent distractor *ba* (which had a VOT of 0 ms). Analysis of the VOTs of the *ta* and *da* utterances from that experiment, however, yielded no significant differences based on whether the distractor was *pa* or *ba*, and also no significant differences between whether the response was *ta* or *ka*. There are reasons to believe that the lack support for this specific prediction in these data is insufficient reason to abandon the model. These reasons are discussed in the next sub-section.

#### 7.3.4. Remaining questions

In DFT, it is equally possible to shorten as well as lengthen a target VOT value by introducing a distractor that has a VOT value that is either somewhat shorter or longer, respectively, than the target utterance. However, Nielsen (2007) showed that there is an asymmetry in the ways in which VOT can be modulated in tasks similar to this one. In her experiments (see section 5.2 of Chapter 5), subjects significantly increased the VOT of *p*- and *k*-initial words as a result stimuli with very long VOTs. Crucially, though, subjects did not significantly shorten VOTs when presented with stimuli with very short VOTs. Shortening the VOT of the voiceless stops was constrained by the phonemic category boundary. In the present experiment, subjects were instructed to reply quickly and, as a result, utterance durations and VOTs were on the shorter range of normal productions due to the speech rate. This means that the VOTs that subjects were producing were already effectively at the floor value as imposed by the phoneme category boundary. Assuming the same constraint on production that Nielsen found here was operative in this experiment, there was simply no “room” for subjects to shorten VOTs further. It would be possible to test the prediction of the model with a modified design of experiment 1 to have *ga-ka* distractors and responses in *ba-pa* and *da-ta* blocks. Two *ka* distractors could be included, one with especially long VOT and one with more prototypical VOT. There should be no constraint in this design preventing the lengthening of *pa* responses on trials with *ka* distractors.

The constraint above and lack of effect in the data from experiment 1 point to an unresolved issue in using a DFT-based model of phonological planning. DFT does not provide a mechanism for enforcing this constraint introduced by the linguistic category. The interaction of the DFT prediction with the linguistic constraints remains an open question for further research and experimentation.

Another area for further potential refinement of DFT-based phonological model to be explored is the appropriate way to address interactions between fields that are described in the model in Chapter 6 as independent, i.e., articulator and voicing. VOT for voiceless stops has been shown to vary with place of articulation (Fischer-Jørgensen, 1954; Peterson & Lehiste, 1960). The model as presented in Chapter 6 does not include a formal mechanism for connecting these fields. It remains a question for future study as to whether that interrelation is best modeled through explicit dependencies between these fields, or at the level of the representations that enter into the planning model. On the other hand, it may not be necessary to modify the model at all if it is simply a physiological artifact arising from factors external to planning, as proposed by Docherty (1992).

It also remains to be determined how and whether the DFT model needs to explicitly address constraints on a parameter like VOT that are imposed by limitations of human physiology and/or perception. For example, the number of categories that can be defined along the VOT continuum while theoretically large seem to be limited to a maximum of three in a given language (Cho & Ladefoged, 1999, and references

therein). Keating (1984) argues that this restriction is a constraint introduced by discriminability in perception, while Löfqvist (1992) proposes that a set of production factors including aerodynamics restrict the speaker's ability to finely control VOT. As with the question of place-dependent VOTs, it is still not clear whether the model of phonological planning is the appropriate place to address these psychological/physiological constraints. The answer to this question may have bearing on the nature of the width of the interaction kernel ( $\sigma_w$ , see section 5.3.2.5 of Chapter 5) and/or on constraints on the shapes of representations of categories on a given continuum.

A few words are also in order regarding the fact that there are other cues to voicing for initial stops in English besides VOT (Kingston & Diehl, 1994; Repp, 1982; Summerfield & Haggard, 1977). Kingston and Diehl (1994) note that these other clues include burst amplitude, fundamental frequency, and first formant transitions. They also argue that at least fundamental frequency modulations are not a reflex of the same mechanism that results in modulations to VOT. In other words, VOT and fundamental frequency are controlled independently. It would be a reasonably straightforward extension of the model to specify all of the particular articulatory settings that are individually controlled, assuming that they are all part of a set of parameters and values that are jointly associated with the feature value in question (e.g., [ $\pm$ voice]). The model is less well equipped at present to accommodate dependencies across fields in this case, similar to the case of articulator-dependent VOT above, if such dependencies were deemed necessary.

#### **7.4. Advantages of dynamical modeling**

The field of phonology has seen a steady increase in research that incorporates experimental methodologies and formal modeling (Coetzee, Kager, & Pater, 2009).

Nevertheless, while response time data have been at the heart of much research in psychology for decades (Luce, 1986), they are rarely used to investigate questions of phonology. One benefit of using a dynamical system to model the process of phonological planning is that this approach enables establishing formal links between phonological processes and response time data, because time is formally incorporated.

In addition to response time data, this dissertation has shown that the use of DFT enables formal models to account for multiple types of data at the same time. For example, in Chapter 5, a single DFT-based model of a classic phoneme categorization task accounted for both gradient response time data and categorical classification data. As just discussed above, the DFT framework also predicts phonetic modulation in the response-distractor task used in Chapter 3 under certain conditions. The use of the DFT framework can therefore account for the same type of data across different tasks, e.g., response times in phoneme classification and the response-distractor tasks, and different types of data within the same task, e.g., response times and categorization in the phoneme classification task, or response times and phonetic modulations within the response-distractor task. This property of this type of model is very desirable because it provides coherence among aspects of experimental results that might need

separate theoretical treatment otherwise. Methodologically, the predictions of such models are falsifiable by more than one kind of data, which makes them better models.

The models presented in Chapters 5 and 6 both make use of inputs that are based on a speaker's representation of the parameter values corresponding to a particular phonological category. These representations are assumed to be homomorphic with the inputs to the model in Chapter 6, that is, phonologically categories are represented as distributions of parameter values on phonetic continua. There is a line of work that has explored using similar representations to account for a variety of phonological phenomena. For example, Gafos and Kirov (Gafos & Kirov, 2010; Kirov & Gafos, 2007) argue that such representations are themselves dynamic, evolving over longer periods of time. They have modeled diachronic lenition as arising from long-term changes in constriction degree of stops as a result of dynamic changes in the representation of that parameter over time due to feedback from usage. Tobin and Nam (2010) use DFT-inspired representations to account for gestural drift (in the sense used by Sancier & Fowler, 1997) in bilingual Spanish-English speakers, who they found modulate the amount of VOT for voiceless stops depending on the amount of exposure to a given ambient language, and the recency of that exposure. Some of the speakers they studied produced longer VOTs in Spanish after recent exposure to English, and shorter VOTs in English after recent exposure to Spanish. According to their account, a category like [-voiced] exists as a single phonological category for these speakers but has two different representations associated with it,

one indexed to Spanish with shorter VOTs and one indexed to English with longer VOTs. The VOT of a given production in either language will be a mixture of the two VOTs, effectively weighted by the recency and amount of exposure to each language. The model of phonological planning presented in Chapter 6 is fully compatible with the notion of dynamical representations, and is ideally suited to implementing “blended” outputs based on multiple inputs. The model is also well-equipped to account for phonetic adjustments in output that speakers seem to plan (e.g., Kingston & Diehl, 1994; Ohala, 1981; Whalen, 1990).

Models using dynamical processes and distributionally defined representations allow for the consideration of new kinds of data in seeking answers about a variety of phonological and psycholinguistic processes. They also show promise for formalizing the interaction between categorical phonological and gradient phonetic phenomena.

## 7.5. Conclusion

During speech production, a person must retrieve the phonological representations of the required utterances by assembling a set of parameter values that specify the vocal tract actions corresponding to these utterances. This dissertation presents a formal computational model of this process. Specifically the model formalizes the process of selecting the particular phonological parameter values for voicing and primary oral articulator needed to eventually produce the required response in a response-distractor task. In the model, assigning values to these

parameters is a time-dependent process, captured as the evolution of a dynamical system over time. The model was developed in part to account for the experimental results presented here, which provide the first clear evidence of independent perceptuo-motor effects of both voicing and articulator. An explicit model of the perception-production link in speech has been lacking in the literature, despite a longstanding debate on the issue. In the present model, the link between perception and production consists of the phonological parameter values of a perceived stimulus obligatorily contributing to the evolution of the activation levels of the fields engaged with the ongoing phonological planning of a required response. The experimental evidence from the present and other studies shows that the contribution of a perceived distractor stimulus can be excitatory when the parameter values of the required response and the distractor are very similar or the same, or can be inhibitory when they are sufficiently different. The present model can explain effects on response times found in the present and other studies, without requiring a special status for complete identity.

Beyond accounting for a variety of experimental results, the model generates predictions that can shed additional light on the process of phonological planning and the interaction between speech perception and production. The experimental and modeling results presented provide greater coherence to the field of psycholinguistics by showing that the fundamental properties involved in the phonological description of linguistic contrast (voicing and articulator) are also parameters that should have a

role in models of speech production. These properties must also be said to be actively involved in the link between perception and production.

## APPENDIX A

MATLAB code (categorize\_trial.m) implementing the simulation of one trial of the phoneme-classification task using the computational model described in Chapter 5.

```
function production = categorize_trial(maxtime,stim_val,plot_what)

%Function calculates Dynamic Field evolution for VOT values based on
% a category pre-shapes and an auditory stimulus, used in the
% phoneme-classification task.

%Parameters:
% - maxtime limits how long the field is allowed to evolve.
% - stim_val is the VOT value of the stimulus, in ms
% - plot_what indicates what to graph:
%   'zip' = don't plot anything (or leave this param empty)
%   'comp' = only the 2 field max activation comparison
%   'all' = max activation levels of the two fields at the end
%           of the evolution, and evolutions of the DA and TA fields
%
% Author: Kevin Roon

%settings: field dimension definition.
% VOT values are defined on an arbitrary scale of -10 to 10.
% (see section 5.3.2.1 of Chapter 5 for conversion to ms)
ll = -10;          % left limit value
rl = 10;           % right limit value
N = 221;           % number of points along each axis
dx = (rl-ll)/(N-1); % the vector that will contain the changes to the
                     % field on one step of the evolution

%settings: within-field parameters
fieldtau = 160;    % time scale ( $\tau$ ) of the evolution of the field
h = -3.25;          % resting state
noisescale = 7;     % noise scaler during convolution ( $weight_{noise}$ )
criterion = 5;      % value ( $\kappa$ ) for selecting a winning field

%settings: interaction kernel parameters
thresh = .7;        % 'soft' threshold ( $\theta$ ) for the interaction term
beta = 1.5;          % slope of the sigmoid function ( $\beta$ )
wexcite = .5;        % local activation strength ( $w_{excite}$ )
winhibit = .15;       % global inhibition strength ( $w_{inhibit}$ )
sigma = 1;           % width of excitatory kernel ( $\sigma_w$ )
ixp = 2*ll:dx:2*rl;
interaction = wexcite.*exp(-(ixp).^2./(2*sigma^2))-winhibit;
```

```

%settings: timing of various events
pre_start = 1;          % starting time of pre-shape input to field
pre_dur = maxtime;       % duration of pre-shape input to field
stim_start = 100;        % starting time of stimulus input to field
stim_dur = 225;          % duration of stimulus input to field

% Define the input for the VOT value for the two fields and
% set up the necessary category pre-shapes

da_val = -4;             % mean of the da distribution ( $val_{DA}$ )
da_strength = .5;         % height of the da distribution ( $p_{DA}$ )
da_width = 1.45;          % SD of the da distribution ( $\sigma_{DA}$ )
S_da = da_strength*S(xp,da_val,da_width);
max_S_da_act = max(S_da);
VOT_S_da = find(S_da==max(S_da), 1,'first');
S_da(1:VOT_S_da) = max_S_da_act;

ta_val = 2;               % mean of the ta distribution ( $val_{TA}$ )
ta_strength = .6;          % height of the ta distribution ( $p_{TA}$ )
ta_width = 1.65;           % SD of the ta distribution ( $\sigma_{TA}$ )
S_ta = ta_strength*S(xp,ta_val,ta_width);
max_S_ta_act = max(S_ta);
VOT_S_ta = find(S_ta==max(S_ta), 1,'first');
S_ta(VOT_S_ta:N) = max_S_ta_act;

% Define the input for the VOT value based on the acoustic distractor
if strcmpi(stim_val,'none');
    stim_weight = 0;      % weight stimulus at 0 if there is none
else
    stim_weight = 2.65;  % weighting of stimulus ( $s$ )
end

stim_width = 1.25;         % width of the stimulus distribution ( $p$ )
if strcmpi(stim_val,'none');
    conv_stim_val = 0;   % assign a meaningless numeric value that
else                      % will be weighted 0 if there is no stimulus
    conv_stim_val = stim_val;  % ( $val_{stimulus}$ )
end
S_stim = stim_weight*S(xp,conv_stim_val,stim_width);

% set up initial state of the two VOT dynamic fields for DA and TA.
% The "adjust" values simply start the field off at the level where
% they will ultimately stabilize if they started at 0. This
% adjustment has no material effect on the results of the model and
% is for visualization purposes only.
pre_adjust = -2;
pre_weight = 2;
yp_da = pre_weight*S_da + pre_adjust;
yp_ta = pre_weight*S_ta + pre_adjust;

```

```

% Set up the fields that will store the data needed for plotting.
% But don't plot anything ('zip') unless explicitly called for
if nargin < 3
    plot_what = 'zip';
end;
if ~strcmpi(plot_what,'zip')
    clf;
    plot_yp_max = zeros(2, maxtime);
end
if strcmpi(plot_what,'all')
    plot_yp_DA = zeros(N,maxtime);
    plot_yp_TA = zeros(N,maxtime);
end;

% Set up variables to keep track of the winner
reach_crit = 0;          % did the model settle on a winner?
crit_i = 0;                % at what time did the model determine a winner
winner = -1;                % did the model choose 'da' (100) or 'ta' (0) or
                            % neither (-1)
VOT_out = -100;           % the max VOT value of the winning field
act_level = -1;           % the activity level of the winning VOT value

% main time loop that builds activation yp values by convolving the
% yp with interaction kernel and input fields as they are received.

for i = 1:maxtime
    S_da_input = zeros(1,N);
    S_ta_input = zeros(1,N);
    % add activation based on category pre-shapes for the duration
    % specified by pre_dur
    if i >= pre_start;
        if i < pre_dur;
            S_da_input = S_da_input + S_da;
            S_ta_input = S_ta_input + S_ta;
        end;
    end

    % add activation to both fields based on acoustic stimulus for
    % the duration specified by stim_dur
    if i >= stim_start;
        if i < stim_start + stim_dur;
            S_da_input = S_da_input + S_stim;
            S_ta_input = S_ta_input + S_stim;
        end;
    end;

    % one time step of letting the field evolve
    dY_da = (1 / fieldtau) * (-yp_da + S_da_input +
        convn(sigmf(yp_da,beta,thresh),interaction,'same')) + h +
        rand([1, N])*(noisescale*rand));

```

```

yp_da = yp_da + dY_da;

dY_ta = (1 / fieldtau) * (-yp_ta + S_ta_input +
    convn(sigmf(yp_ta,beta,thresh),interaction,'same') + h +
    rand([1, N])*(noisescale*rand));
yp_ta = yp_ta + dY_ta;

% at the end of each i step, check to see whether the maximum
% activation level of either field has crossed the criterion
% value. If so, choose that field as the winning categorization,
% make this i the effective RT, record activation and VOT values,
% exit the for loop. Else, go until maxtime.

max_da = max(yp_da);
max_ta = max(yp_ta);
if reach_crit == 0;
    if max_da >= criterion
        winner = 100;
        crit_i = i;
        reach_crit = 1;
        VOT_out = ll + (find(yp_da==max(yp_da), 1,'first')*dx);
        act_level = max_da;
    end;
    if max_ta >= criterion
        winner = 0;
        crit_i = i;
        reach_crit = 1;
        VOT_out = ll + (find(yp_ta==max(yp_ta), 1,'first')*dx);
        act_level = max_ta;
    end;
end;
if ~strcmpi(plot_what,'zip')
    plot_yp_max(1,i) = max_da;
    plot_yp_max(2,i) = max_ta;
end

if strcmpi(plot_what,'all')
    plot_yp_DA(:,i) = yp_da;
    plot_yp_TA(:,i) = yp_ta;
end;
end;

VOT_da = ll + (find(yp_da==max(yp_da), 1,'first')*dx);
VOT_ta = ll + (find(yp_ta==max(yp_ta), 1,'first')*dx);

if ~strcmpi(plot_what,'zip')
    % Plot the comparison of maximum activations evolutions of the DA
    % and TA fields
    crit_line = zeros(crit_i) + criterion;
    within_line = zeros(maxtime) + thresh;

```

```

p = plot(plot_yp_max(1,:), '--b');
set(p,'LineWidth',1.1)
hold on
p = plot(plot_yp_max(2,:), 'r');
set(p,'LineWidth',1.1)
legend('DA field','TA field');
title('Maximum activation levels','FontSize',15,
'FontWeight','b');
set(gca,'FontSize',15)
xlabel('time steps');
ylabel('activation');
plot(crit_line, 'k');
plot(within_line,'-.k');
set(gca,'yLim',[-2 6]);
y_min = min(get(gca, 'yLim'));
y_max = max(get(gca, 'yLim'));
plot([crit_i, crit_i],[y_min, criterion], 'k');
if ~strcmp(stim_val, 'none')
    plot([stim_start, stim_start],[y_min, y_max], ':k');
    plot([stim_start+stim_dur, stim_start+stim_dur],[y_min,
y_max], ':k');
end
end
if strcmp(plot_what,'all')
    yTickVals = [0 45 78 89 100 111 122 155 177 225];
    yTickLabels = [-70 -30 0 10 20 30 40 70 90 130];
    %Plot the detailed evolution of the DA field
    figure
    meshc(plot_yp_DA);
    hold on;
    axis ij;
    colormap(copper);
    title('DA field evolution','FontSize',14, 'FontWeight','b');
    set(gca,'FontSize',14)
    ylabel('VOT (ms)')
    xlabel('time')
    zlabel('activation','Rotation',90);
    set(gca,'yLim',[0 225]);
    set(gca,['y' 'Tick'], yTickVals);
    set(gca,['y' 'TickLabel'], yTickLabels);
    set(gca,'zLim',[-8 6]);

    %Plot the detailed evolution of the TA field
    figure
    meshc(plot_yp_TA);
    hold on;
    axis ij;
    colormap(copper);
    title('TA field evolution','FontSize',14, 'FontWeight','b');
    set(gca,'FontSize',14)

```

```

ylabel('VOT (ms)')
xlabel('time')
zlabel('activation','Rotation',90);
set(gca,'yLim',[0 225]);
set(gca,['y' 'Tick'], yTickVals);
set(gca,['y' 'TickLabel'], yTickLabels);
set(gca,'zLim',[-8 6]);
end;

crit_i = crit_i - stim_start;

production = [reach_crit, winner, crit_i, VOT_out, act_level, VOT_da,
              VOT_ta];

function input = S(x,off,stddev)
%input function
input = exp(-((1*(x-off)).^2)/(2*stddev^2));

function x = sigmf(y, beta, thresh)
%sigmoid function, see (5) in section 5.3.2.5 of Chapter 5
x = 1./(1+exp(-beta.*(y-thresh)));

```

## APPENDIX B

MATLAB code (categorize\_exp.m) for implementing simulations of the phoneme-classification experiment, i.e., multiple trials for any number of stimuli, as described in Chapter 5.

```
function guess_per_cat = categorize_exp(stimuli, repetitions)
% 'stimuli' is a vector (any size) that contains the stimuli values
% to be simulated.
% The number of simulations per stimulus value are specified by
% 'repetitions'
% guess_per_cat returns a vector that indicates the number of trials
% per stimulus category where the categorization did not pick a
% winner, and was instead assigned by chance. [this was 0 in the
% model simulations reported in Chapter 5, i.e., it did not happen]

% Author: Kevin Roon

clf;
num_of_stim = size(stimuli, 2);
trial_log = zeros(repetitions, num_of_stim);

for i = 1:num_of_stim;
    stim_val = stimuli(1, i);
    for j = 1:repetitions;
        trial_result = decide_trial(stim_val);
        trial_log (j,i,1) = trial_result(1);
        trial_log (j,i,2) = trial_result(2);
        trial_log (j,i,3) = trial_result(3);
    end;
end;

X_axis = (stimuli + 3) * 10;
RT_means = mean(trial_log(:,:,2));
RT_error = std(trial_log(:,:,2));
% RT_error = zeros(1, num_of_stim);
errorbar(stimuli, RT_means, RT_error, 'Marker', 'd','LineWidth',1.1);
title('Simulated response times','FontSize',15, 'FontWeight','b');
set(gca,'FontSize',15)
xlim([min(stimuli)-1 max(stimuli)+1]);
xlabel('VOT of stimulus');
ylabel('time steps');
set(gca, 'XTick', stimuli);
set(gca, 'XTickLabel', X_axis);
```

```

figure
ID_means = mean(trial_log(:,:,1));
ID_error = zeros(1, num_of_stim);
%ID_error = std(trial_logID); %I don't think this is right.
errorbar(stimuli, ID_means, ID_error, 'Marker', 'o','LineWidth',1.1);
title('Simulated classifications','FontSize',15, 'FontWeight','b');
set(gca,'FontSize',15)
xlim([min(stimuli)-1 max(stimuli)+1]);
xlabel('VOT of stimulus');
ylim([-10 110]);
ylabel('% "da" responses');
set(gca, 'XTick', stimuli);
set(gca, 'XTickLabel', X_axis);

guess_per_cat = trial_log;

function trial_result = decide_trial(vot)
maxtime = 2000;
trial = categorize_trial(maxtime,vot);
guessed = 0;
if trial(1) == 1;
    response = trial(2);
    response_time = trial(3);
else %choose at random
    guessed = guessed + 1;
    response_time = maxtime;
    x = rand;
    if x > 0.5;
        response = 100;
    else
        response = 0;
    end;
end;
trial_result = [response response_time guessed];

```

## APPENDIX C

MATLAB code (dft\_resp\_distr.m) implementing the simulation of one trial of the response-distractor task using the computational model described in Chapter 6.

```
function production = dft_resp_distr(option1, option2, resp, distr,
SOA, maxtime, plot_what)

% Function calculates the Dynamic Field evolutions for a single trial
% of the response-distractor task. The output of the field is a
% vector of the final values of the field, with some other data (see
% definition at end).
% Parameters:
% option 1: In a given block, 1 of the 2 possible responses.
% option 2: The other of the possible responses in that block.
% resp: The required response. One of
% {pa, ta, da, ka, ga, ma, na, 'none'}.
% 'none' shows how the fields behave when no
% response is required.
% distr: The perceived distractor. One of {pa, ba, da, ga, tone}.
% Pass 'tone' to have no linguistic distractor influence.
% SOA: The time between the onsets of the presentation of the
% visual cue indicating the required response and the
% distractor.
% 100, 200, 300 are the options in the experiment, but it
% can be set to anything.
% maxtime: Limits how long the field is allowed to evolve.
% plot_what: Indicates whether and what the script should plot.
% Possible:
% 0 = don't plot anything (default if not set)
% Set it to 0 if this function is being called by another.
% 1 = comparative field max evolution only.
% 2 = comparative field and the detailed evolutions of the
% four activation fields

% Author: Kevin Roon, Spring 2012

% settings: field definition
% VOT and Constriction Location (CL) values are defined on an
% arbitrary scale of -10 to 10.
ll = -10; % left limit
rl = 10; % right limit
N = 221; % landscape detail = number of points along
% x axis
dx = (rl-ll)/(N-1);
xp = ll:dx:rl;
```

```

% settings: within-field parameters
fieldtau = 150; % field evolution time scale ( $\tau$ )
h = -3.25; % resting level of all fields
noisescale = 5; % noise scaler during convolution

% settings: within-field interaction kernel parameters
thresh = .7; % 'soft' interaction threshold ( $\theta$ )
wexcite = .45; % local activation strength ( $w_{excite}$ )
winhibit = .1; % global inhibition strength ( $w_{inhibit}$ )
sigma = 1; % width of excitatory kernel ( $\sigma_w$ )
beta = 1.5; % the slope of the sigmoid function for
% determining which values enter into the
% interaction kernel ( $\beta$ )

ixp = 2*11:dx:2*rl;
interaction = wexcite.*exp(-(ixp).^2./(2*sigma^2))-winhibit;

% settings: timing of various events
pre_start = 1; % "pre"-shapes persist through the whole
% evolution
pre_dur = maxtime; % how long should the pre-shape be present
resp_start = 100; % indicates when response input should start
resp_dur = 400; % duration of the response-based inputs
distr_start = resp_start + SOA; % by definition
distr_dur = 125; % duration of the distractor inputs

% settings: cross-field inhibition
cross_thresh = 0; % value an articulator field activation level
% has to cross to inhibit other fields ( $\chi$ )
cross_inhibit = 1.25; % amount by which inhibiting field reduces
% activation levels of other fields, ( $q$ )

% weights of the various input types
pre_weight = .8; % weight of the pre-shape inputs ( $p$ )
if strcmpi(resp, 'none');
    resp_weight = 0; % no response required, so weight it at 0
else
    resp_weight = 2.7; % weighting factor of response ( $r$ )
end

if strcmpi(distr, 'tone');
    distr_weight_art = 0; % no linguistic distractor, so set those
    distr_weight_vc = 0; % weights to 0
else
    distr_weight_art = 7.5; % weighting of place feature for
    % distractor ( $d_{artic}$ )
    distr_weight_vc = 6.3; % weighting of voice feature for
    % distractor ( $d_{voice}$ )
end

```

```

% settings: Monitor parameters
criterion = 6;      % Criterion for determining production value ( $\kappa$ )
winner = 0;          % Has the monitor chosen a winning art-VOT combo?
crit_art = 0;        % Which articulator field hit criterion first?
art_CL_out = 0;      % What was the max CL value for that articulator?
crit_i_art = 0       % Time step at which some articulator value
                     % hit criterion.
crit_i_VOT = 0;      % Time step at which some VOT value hit criterion.
VOT_out = -1;         % VOT value with max activation when winner chosen.

% Set up initial state of the VOT dynamic field (yp_VOT) and the
% three articulator fields (yp_LL, yp_TT, yp_TB)
% The pre_adjust, pre_adjust_wt, and pre_adjust_VOT values are there
% to start the fields roughly at the place that they stabilize
% without any input other than the pre-shapes. Simply a time-saving
% measure, as the field will soon stabilize to the values close to
% these if not adjusted. Has no material effect on the outcome.
% The specific values for these depend on the values of h,
% noisyscale, etc.
pre_adjust = -2;
pre_adjust_wt = 1.1;
pre_adjust_VOT = .75;
yp_LL = zeros(1,N) + pre_adjust;
yp_TT = zeros(1,N) + pre_adjust;
yp_TB = zeros(1,N) + pre_adjust;
yp_VOT = zeros(1,N) + pre_adjust;

% DEFINE ALL INPUTS to the fields
% First, PRE-SHAPES for the 4 fields, starting with articulators...

% Get all the parameter values for the two possible responses
option1_vals = input(option1);
option2_vals = input(option2);

% Set up the pre-shape for the LL field if there are potential labial
% responses. First check the first possible response.
S_option1LL = 0;
S_option2LL = 0;
if option1_vals(:, 1) == 1;
    S_option1LL = S(xp, option1_vals(:, 2), option1_vals(:, 3));
end;

% Then check the second potential response.
if option2_vals(:, 1) == 1;
    S_option2LL = S(xp, option2_vals(:, 2), option2_vals(:, 3));
end;

```

```

% Check whether the pre-shape input to the field for option1 and
% option1 have the same articulator. If so, don't double the input.
if option1_vals(:, 1) == option2_vals(:, 1);
    S_preLL = S_option1LL;
else
    S_preLL = S_option1LL + S_option2LL;
end;

% Adjust the starting state of LL field to get it to where it
% stabilizes without response or distractor input if there is
% LL pre-shape.
if option1_vals(:, 1) == 1 || option2_vals(:, 1) == 1;
    yp_LL = yp_LL + S_preLL*pre_adjust_wt;
end

% Now do the same for the TT field...
S_option1TT = 0;
S_option2TT = 0;
if option1_vals(:, 1) == 2;
    S_option1TT = S(xp, option1_vals(:, 2), option1_vals(:, 3));
end;
if option2_vals(:, 1) == 2;
    S_option2TT = S(xp, option2_vals(:, 2), option2_vals(:, 3));
end;
if option1_vals(:, 1) == option2_vals(:, 1);
    S_preTT = S_option1TT;
else
    S_preTT = S_option1TT + S_option2TT;
end;

if option1_vals(:, 1) == 2 || option2_vals(:, 1) == 2;
    yp_TT = yp_TT + S_preTT*pre_adjust_wt;
end

% ...and for the TB field...
S_option1TB = 0;
S_option2TB = 0;
if option1_vals(:, 1) == 3;
    S_option1TB = S(xp, option1_vals(:, 2), option1_vals(:, 3));
end;
if option2_vals(:, 1) == 3;
    S_option2TB = S(xp, option2_vals(:, 2), option2_vals(:, 3));
end;
if option1_vals(:, 1) == option2_vals(:, 1);
    S_preTB = S_option1TB;
else
    S_preTB = S_option1TB + S_option2TB;
end;

```

```

if option1_vals(:, 1) == 3 || option2_vals(:, 1) == 3;
    yp_TB = yp_TB + S_preTB*pre_adjust_wt;
end

% ...and lastly, the VOT field. Here, two different inputs are
% possible if the potential responses differ in voicing.
if option1_vals(:, 4) == option2_vals(:, 4);
    S_prevVOT = S(xp, option1_vals(:, 4), option1_vals(:, 5));
    yp_VOT = yp_VOT + S_prevVOT*pre_adjust_wt;
else
    S_prevVOT = S(xp, option1_vals(:, 4), option1_vals(:, 5)) + S(xp,
option2_vals(:, 4), option2_vals(:, 5));
    yp_VOT = yp_VOT + S_prevVOT*pre_adjust_VOT;
end;

% Second, the required RESPONSE input
resp_vals = input(resp);
% Add a little noise to the CL and VOT values on this trial. This
% implies that there are various factors not included in the model
% that result in a speaker's productions varying a little from
% utterance to utterance.
resp_vals(:, 2) = resp_vals(:, 2)+(rand-.5);
resp_vals(:, 4) = resp_vals(:, 4)+(rand-.5);

% Define the inputs to the three articulator fields...
S_respLL = 0;
S_respTT = 0;
S_respTB = 0;
if resp_vals(:, 1) == 1;
    S_respLL = S(xp, resp_vals(:, 2), resp_vals(:, 3));
end;
if resp_vals(:, 1) == 2;
    S_respTT = S(xp, resp_vals(:, 2), resp_vals(:, 3));
end;
if resp_vals(:, 1) == 3;
    S_respTB = S(xp, resp_vals(:, 2), resp_vals(:, 3));
end;

% ...and then the VOT field.
S_respVOT = S(xp, resp_vals(:, 4), resp_vals(:, 5));

% Last, the define the inputs based on the DISTRACTOR
S_distrLL = 0;
S_distrTT = 0;
S_distrTB = 0;
S_distrVOT = 0;
if ~strcmpi(distr, 'tone');
% Don't do anything if the distractor is a tone
    distr_vals = input(distr);
% Otherwise, set up the inputs for the articulator fields...

```

```

if distr_vals(:, 1) == 1;
    S_distrLL = S(xp, distr_vals(:, 2), distr_vals(:, 3));
end;
if distr_vals(:, 1) == 2;
    S_distrTT = S(xp, distr_vals(:, 2), distr_vals(:, 3));
end;
if distr_vals(:, 1) == 3;
    S_distrTB = S(xp, distr_vals(:, 2), distr_vals(:, 3));
end;
% ...and the VOT field.
S_distrVOT = S(xp, distr_vals(:, 4), distr_vals(:, 5));
end

% Set up arrays for plotting, if needed
if nargin < 7;
    plot_what = 0;
end;
if plot_what > 0;
    clf;
    plot_yp_max = zeros(4,maxtime);
    if plot_what > 1;
        plot_yp_LL = zeros(N,maxtime);
        plot_yp_TT = zeros(N,maxtime);
        plot_yp_TB = zeros(N,maxtime);
        plot_yp_VOT = zeros(N,maxtime);
    end
end;

% MAIN LOOP
for i = 1:maxtime
% Calculate the inputs to be added to the fields on each timestep,
    S_LL_inputs = zeros(1,N);
    S_TT_inputs = zeros(1,N);
    S_TB_inputs = zeros(1,N);
    S_VOT_inputs = zeros(1,N);

% and the amount of cross-field inhibition needed on each i.
    S_cross_LL = zeros(1,N);
    S_cross_TT = zeros(1,N);
    S_cross_TB = zeros(1,N);

% Add pre-shape activation
    if i >= pre_start;
        if i < pre_start + pre_dur;
            S_LL_inputs = S_LL_inputs + S_preLL*pre_weight;
            S_TT_inputs = S_TT_inputs + S_preTT*pre_weight;
            S_TB_inputs = S_TB_inputs + S_preTB*pre_weight;
            S_VOT_inputs = S_VOT_inputs + S_preVOT*pre_weight;
        end
    end
end

```

```

% Add activation based on the required response
if i >= resp_start;
    if i < resp_start + resp_dur;
        S_LL_inputs = S_LL_inputs + S_respLL*resp_weight;
        S_TT_inputs = S_TT_inputs + S_respTT*resp_weight;
        S_TB_inputs = S_TB_inputs + S_respTB*resp_weight;
        S_VOT_inputs = S_VOT_inputs + S_respVOT*resp_weight;
    end
end

% Add activation based on the acoustic distractor
if i >= distr_start;
    if i < distr_start + distr_dur;
        S_LL_inputs = S_LL_inputs + S_distrLL*distr_weight_art;
        S_TT_inputs = S_TT_inputs + S_distrTT*distr_weight_art;
        S_TB_inputs = S_TB_inputs + S_distrTB*distr_weight_art;
        S_VOT_inputs = S_VOT_inputs + S_distrVOT*distr_weight_vc;
    end;
end;

% Calculate inhibition from competing fields. Determine which if
% any articulator fields have exceeded threshold. For each that
% has, subtract activation from the other 2 articulator fields.
max_LL = max(yp_LL);
max_TT = max(yp_TT);
max_TB = max(yp_TB);
if max_LL >= cross_thresh;
    S_cross_TT = S_cross_TT + cross_inhibit;
    S_cross_TB = S_cross_TB + cross_inhibit;
end;
if max_TT >= cross_thresh;
    S_cross_LL = S_cross_LL + cross_inhibit;
    S_cross_TB = S_cross_TB + cross_inhibit;
end;
if max_TB >= cross_thresh;
    S_cross_LL = S_cross_LL + cross_inhibit;
    S_cross_TT = S_cross_TT + cross_inhibit;
end;
if plot_what > 0; % Update the plotting arrays, if required
    plot_yp_max(1,i) = max_LL;
    plot_yp_max(2,i) = max_TT;
    plot_yp_max(3,i) = max_TB;
    plot_yp_max(4,i) = max(yp_VOT);
    if plot_what > 1;
        plot_yp_LL(:,i) = yp_LL;
        plot_yp_TT(:,i) = yp_TT;
        plot_yp_TB(:,i) = yp_TB;
        plot_yp_VOT(:,i) = yp_VOT;
    end
end;

```

```

% THIS IS WHERE THE DFT WORK IS DONE: one time step of letting
% the fields evolve. First the articulators...
dY_LL = (1 / fieldtau) * (-yp_LL + S_LL_inputs - S_cross_LL +
convn(sigmf(yp_LL, beta, thresh),interaction,'same') + h + rand([1,
N])*noisescale*rand));
yp_LL = yp_LL + dY_LL;

dY_TT = (1 / fieldtau) * (-yp_TT + S_TT_inputs - S_cross_TT +
convn(sigmf(yp_TT, beta, thresh),interaction,'same') + h + rand([1,
N])*noisescale*rand));
yp_TT = yp_TT + dY_TT;

dY_TB = (1 / fieldtau) * (-yp_TB + S_TB_inputs - S_cross_TB +
convn(sigmf(yp_TB, beta, thresh),interaction,'same') + h + rand([1,
N])*noisescale*rand));
yp_TB = yp_TB + dY_TB;

% ...then VOT.
dY_VOT = (1 / fieldtau) * (-yp_VOT + S_VOT_inputs +
convn(sigmf(yp_VOT,beta,thresh),interaction,'same') + h + rand([1,
N])*noisescale*rand));
yp_VOT = yp_VOT + dY_VOT;

% MONITOR
% Check to see whether some value in both the VOT field and one
% articulator field have passed criterion, unless there was a
% winner chosen already.
if winner == 0;
    if crit_i_VOT == 0;
        if max(yp_VOT) >= criterion;
            crit_i_VOT = i;
        end;
    end;
    if crit_i_art == 0;
        if max(yp_LL) >= criterion;
            crit_art = 1;
            crit_i_art = i;
            art_CL_out = (find(yp_LL==max(yp_LL), 1,'first')*dx);
        end;
        if max(yp_TT) >= criterion;
            crit_art = 2;
            crit_i_art = i;
            art_CL_out = (find(yp_TT==max(yp_TT), 1,'first')*dx);
        end;
        if max(yp_TB) >= criterion;
            crit_art = 3;
            crit_i_art = i;
            art_CL_out = (find(yp_TB==max(yp_TB), 1,'first')*dx);
        end;
    end;
end;

```

```

        if crit_i_VOT && crit_i_art;
            VOT_out = ll + (find(yp_VOT==max(yp_VOT), 1,'first')*dx);
            winner = i;
        end
    end
end;

if VOT_out == 0;
    VOT_out = ll + (find(yp_VOT==max(yp_VOT), 1,'first')*dx);
end

if plot_what > 0;
%Plot the comparison of max activations evolutions of all 4 fields
    crit_line = zeros(winner) + criterion;
    cross_line = zeros(maxtime) + cross_thresh;
    within_line = zeros(maxtime) + thresh;

    p = plot(plot_yp_max(1,:), '--b');
    set(p,'LineWidth',1.1)
    hold on
    p = plot(plot_yp_max(2,:), '-.r');
    set(p,'LineWidth',1.1)
    hold on
    p = plot(plot_yp_max(3,:), '-m');
    set(p,'LineWidth',.5)
    hold on
    p = plot(plot_yp_max(4,:), '-k');
    set(p,'LineWidth',1.1)
    plot(crit_line, 'k');
    plot(cross_line,'-.c');
    plot(within_line,'-.k');
    set(gca,'FontSize',16);      %11
    legend('LL', 'TT', 'TB', 'Voicing');
    if ~strcmp(distr, 'tone');
        title({'Comparison of Field Evolutions';[option1 '/' option2
        ' block, response: ' resp ', distractor: ' distr ', SOA =
        int2str(SOA)]; },'FontSize',14, 'fontweight','b');
    else
        title({'Comparison of Field Evolutions';[option1 '/' option2
        ' block, response: ' resp ', distractor: ' distr]; },'FontSize',14,
        'fontweight','b');
    end
    ylabel('activation','Rotation',90);
    xlabel('time')

%Plot the vertical lines.
y_min = min(get(gca, 'yLim'));
y_max = max(get(gca, 'yLim'));

```

```

% Plot the point at which the winning RT was determined,
% if there was one.
plot([winner, winner],[y_min, criterion], 'k');

% Plot the start and end points of the input corresponding
% to the response.
if ~strcmpi(resp, 'none');
    plot([resp_start, resp_start],[y_min, y_max], ':k');
    plot([resp_start+resp_dur, resp_start+resp_dur],[y_min,
y_max], ':k');
end

% Plot the start and end points of the input corresponding
% to the distractor.
if ~strcmpi(distr, 'tone');
    plot([distr_start, distr_start],[y_min, y_max], '--k');
    plot([distr_start+distr_dur, distr_start+distr_dur],[y_min,
y_max], '--k');
end

if plot_what > 1;
    % Indicates that plots of all fields are requested.
    %First, plot the evolution of the LL field.
    figure
    meshc(plot_yp_LL);
    hold on;
    axis ij;
    colormap(copper);
    set(gca,'FontSize',14)
    if ~strcmpi(distr, 'tone');
        title({'LL Field Evolution';[option1 '/' option2 ' block,
response: ' resp ', distractor: ' distr ', SOA = ' int2str(SOA)];
}, 'FontSize',14);
    else
        title({'LL Field Evolution';[option1 '/' option2 ' block,
response: ' resp ', distractor: ' distr ]; }, 'FontSize',14);
    end
    ylabel('Constriction Location')
    xlabel('time')
    zlabel('activation','Rotation',90);
    yTicks = get(gca,['y' 'Tick']);
    set(gca,['y' 'TickLabel'], yTicks);

    ...then plot the evolution of the TT field...
    figure
    meshc(plot_yp_TT);
    hold on;
    axis ij;
    colormap(copper);
    set(gca,'FontSize',14)

```

```

if ~strcmpi(distr, 'tone');
    title({'TT Field evolution';[option1 '/' option2 ' block,
response: ' resp ', distractor: ' distr ', SOA = ' int2str(SOA)];
},'FontSize',14);
else
    title({'TT Field evolution';[option1 '/' option2 ' block,
response: ' resp ', distractor: ' distr]; },'FontSize',14);
end
ylabel('Constriction Location')
xlabel('time')
yTicks = get(gca,['y' 'Tick']);
set(gca,['y' 'TickLabel'], yTicks);
zlabel('activation','Rotation',90);

%...then plot the evolution of the TB field...
figure
meshc(plot_yp_TB);
hold on;
axis ij;
colormap(copper);
set(gca,'FontSize',14)
if ~strcmpi(distr, 'tone');
    title({'TB Field evolution';[option1 '/' option2 ' block,
response: ' resp ', distractor: ' distr ', SOA = ' int2str(SOA)];
},'FontSize',14);
else
    title({'TB Field evolution';[option1 '/' option2 ' block,
response: ' resp ', distractor: ' distr]; },'FontSize',14);
end
ylabel('Constriction Location')
xlabel('time')
yTicks = get(gca,['y' 'Tick']);
set(gca,['y' 'TickLabel'], yTicks);
zlabel('activation','Rotation',90);

%...and lastly, plot the evolution of the VOT field.
figure
meshc(plot_yp_VOT);
hold on;
axis ij;
colormap(bone);
set(gca,'FontSize',14)
if ~strcmpi(distr, 'tone');
    title({'Voicing Field evolution';[option1 '/' option2 ' block,
response: ' resp ', distractor: ' distr ', SOA = ' int2str(SOA)];
},'FontSize',14);
else
    title({'Voicing Field evolution';[option1 '/' option2 ' block,
response: ' resp ', distractor: ' distr]; },'FontSize',14);
end

```

```

        ylabel('VOT (in ms)')
        xlabel('time')
        yTicks = [0, 45, 90, 135, 180, 221];
        set(gca,['y' 'TickLabel'], yTicks);
        zlabel('activation','Rotation',90);
    end
end;
production = [winner-resp_start, crit_art, art_CL_out, crit_i_art,
VOT_out, crit_i_VOT];

function vals = input(syllable)
% Determines the articulator field and its CL value, and the VOT
% field value for pre-shapes, responses, and distractors.
% Can handle {pa, ba, ta, da, ka, ga, ma, na}
LL = 'pa, ba, ma';
TT = 'ta, da, na';
TB = 'ka, ga';
voiced = 'ba, da, ga';
voiceless = 'pa, ta, ka';
nasal = 'ma, na';
articulator = 0;
if ~isempty(strfind(LL, syllable));
    articulator = 1;
elseif ~isempty(strfind(TT, syllable));
    articulator = 2;
elseif ~isempty(strfind(TB, syllable));
    articulator = 3;
end

% Since the CL does not vary within articulator, all of them will be
% set to 0, and the CL_width will be the same for all fields.
CL = 0;
CL_width = 2;

% Specify VOT in terms of the -10:10 scale of the field. Will be
% converted to ms values automatically on the graph later.
VOT = 0;
VOT_width = 1;
if ~isempty(strfind(voiced, syllable));
    VOT = -5;
    VOT_width = 3;
elseif ~isempty(strfind(voiceless, syllable));
    VOT = 5;
    VOT_width = 3;
elseif ~isempty(strfind(nasal, syllable));
    VOT = -7;
    VOT_width = 2;
end

vals = [articulator, CL, CL_width, VOT, VOT_width];

```

```
function input = s(x,off,stdev)
%input function
input = exp(-((1*(x-off)).^2)/(2*stdev^2));

function x = sigmf(y, beta, thresh)
%sigmoid function
x = 1./(1+exp(-beta.*(y-thresh)));
```

## APPENDIX D

MATLAB code (simulate\_exps.m) implementing the simulation of the response-distractor experiment, i.e., multiple trials and conditions, using the computational model defined in Chapter 6.

```
function simulate_data = simulate_exps(experiment, repetitions, SOA)
% Simulates the response-distractor experiments. Graphs the
% simulated RTs for four conditions: ID, Tone (=No Distractor),
% Congruent, and Incongruent
% Parameters:
% - experiment: 1 or 2, depending on which experiment is simulated,
%   Or, pass 3 to simulate both experiments 1 and 2 and overlay the
%   results.
% - repetitions: indicates the number of trials *per block* (i.e.,
%   actual number of trials will be repetitions * 4 [the
%   number of conditions].
% - "SOA": in time steps (i)

% Author: Kevin Roon, Spring 2012

clf;
if experiment == 3;
    trial_results = simulate(1, repetitions, SOA);
else
    trial_results = simulate(experiment, repetitions, SOA);
end
X_axis = [1 2 3 4];

% Plot the RTs by condition
RT_means = mean(trial_results(:, :, 1));
RT_error = std(trial_results(:, :, 1));
errorbar(X_axis, RT_means, RT_error, 'Marker', 'd', 'Color', 'r',
'LineStyle', '--', 'LineWidth', 1.25);
if experiment ~= 3;
    title({['Simulation of RTs for Experiment ' num2str(experiment)];
    [num2str(repetitions) ' trials per condition. SOA = '
    num2str(SOA)];}, 'FontSize', 15);
end

set(gca, 'FontSize', 15)
xlabel('Distractor Conditions');
set(gca, 'yLim', [150 450]);
ylabel('Response times', 'fontWeight', 'b');
set(gca, 'XTick', X_axis);
set(gca, 'XTickLabel', {'Identity'; 'Tone'; 'Congruent'; 'Incongruent'},
'fontWeight', 'b')
```

```

if experiment == 3;
    trial_results = simulate(2, repetitions, SOA);
    hold on
    RT_means = mean(trial_results(:, :, 1));
    RT_error = std(trial_results(:, :, 1));
    errorbar(X_axis, RT_means, RT_error, 'Marker', 'd', 'Color', 'b',
    'LineStyle', '--', 'LineWidth', 1.25);
    title({'Simulation of RTs for Experiments 1 (solid) and 2
    (dashed)'; [num2str(repetitions) ' trials per condition. SOA = '
    num2str(SOA)]}, 'FontSize', 16);
    set(gca, 'FontSize', 15)
end
simulate_data = trial_results;
end

function trial_log = simulate(exp, reps, SOA)
trial_log = zeros(reps, 4);

if exp == 1; % For experiment 1
    option1 = 'ta'; % Options 1 and 2 indicate the possible responses
    option2 = 'da'; % on a given trial. These define the pre-shapes
    % of the fields
    distr1 = 'pa'; % The distractor in the Congruent condition
    distr2 = 'ba'; % The distractor in the Incongruent condition
elseif exp == 2; % as above, but for experiment 2
    option1 = 'ta';
    option2 = 'ka';
    distr1 = 'da';
    distr2 = 'ga';
end

for i = 1:reps;
    % for option1 responses
    % ID condition (1)
    trial_result = dft_resp_distr(option1, option2, option1, option1,
    SOA, 2000, 0);
    trial_log(i, 1, 1) = trial_result(1);
    trial_log(i, 1, 2) = trial_result(2);
    trial_log(i, 1, 3) = trial_result(5);

    % tone condition (2)
    trial_result = dft_resp_distr(option1, option2, option1, 'tone',
    SOA, 2000, 0);
    trial_log(i, 2, 1) = trial_result(1);
    trial_log(i, 2, 2) = trial_result(2);
    trial_log(i, 2, 3) = trial_result(5);

```

```
% congruent condition (3)
trial_result = dft_resp_distr(option1, option2, option1, distr1,
SOA, 2000, 0);
trial_log(i, 3, 1) = trial_result(1);
trial_log(i, 3, 2) = trial_result(2);
trial_log(i, 3, 3) = trial_result(5);

% incongruent condition (4)
trial_result = dft_resp_distr(option1, option2, option1, distr2,
SOA, 2000, 0);
trial_log(i, 4, 1) = trial_result(1);
trial_log(i, 4, 2) = trial_result(2);
trial_log(i, 4, 3) = trial_result(5);
end
end
```

## REFERENCES

- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (under review). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*.
- Bates, D. M. (2005). Fitting linear mixed models in R. *R News*, 5, 27–30.
- Bates, D. M., & Maechler, M. (2009). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-31.  
<http://CRAN.R-project.org/package=lme4>.
- Belsley, D. A., Kuh, E., & Welsch, R. E. (2004). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley.
- Blumstein, S. E., Myers, E. B., & Rissman, J. (2005). The perception of Voice Onset Time: An fMRI investigation of phonetic category structure. *Journal of Cognitive Neuroscience*, 17(9), 1353–1366.

Boersma, P. (1998). *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. Ph.D., University of Amsterdam, Amsterdam.

Boersma, P., & Weenink, D. (2006). Praat 4.4: Doing phonetics by computer.  
Retrieved from <http://www.praat.org>

Bohland, J. W., Bullock, D., & Guenther, F. H. (2009). Neural representations and mechanisms for the performance of simple speech sequences. *Journal of Cognitive Neuroscience*, 22(7), 1504–1529. doi: 10.1162/jocn.2009.21306

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

Browman, C. P., & Goldstein, L. M. (1988). Some notes on syllable structure in Articulatory Phonology. *Phonetica*, 45, 140–155.

Browman, C. P., & Goldstein, L. M. (1989). Articulatory gestures as phonological units. *Phonology*, 6, 201-251.

Browman, C. P., & Goldstein, L. M. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, 18, 299-320.

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics*, 27, 207-229.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Coetzee, A. W., Kager, R., & Pater, J. (2009). Introduction: Phonological models and experimental data. *Phonology*, 26(1), 1–8.

Cole, J., Kim, H., Choi, H., & Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics*, 35(2), 180–209.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, 26, 55–67.

Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *NeuroReport*, 8(4), 919–924.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93(3), 283–321.

Dell, G. S. (1988). The retrieval of phonological forms in production: tests of predictions from a connectionist model. *Journal of Memory and Language*, 27(2), 124–142.

- Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraint in phonological speech errors. *Cognitive Science*, 17, 149–195.
- Dell, G. S., & O'Seaghda, P. G. (1992). Stages of lexical access in language production. *Cognition*, 42, 287–314.
- Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, 55, 149–179.
- Docherty, G. J. (1992). *The timing of voicing in British English obstruents*. Berlin; New York: Foris.
- Dutilh, G., Vandekerckhove, J., Forstmann, B., Keuleers, E., Brysbaert, M., & Wagenmakers, E.-J. (2012). Testing theories of post-error slowing. *Attention, Perception, & Psychophysics*, 7(4), 454–465.
- Echambadi, R., & Hess, J. D. (2007). Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science*, 26(3), 438–445.
- Eimas, P. D., & Corbit, J. D. (1973). Lective adaptation of linguistic feature detectors. *Cognitive Psychology*, 4(1), 99–109.
- Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, 109(3), 545–572.

Fischer-Jørgensen, E. (1954). Acoustic analysis of stop consonants. *Miscellanea Phonetica*, 2, 197–221.

Fowler, C. A. (1979). “Perceptual centers” in speech production and perception. *Perception & Psychophysics*, 25(5), 375–388.

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

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, 99(3), 1730–1741.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, 49(3), 396–413.

Frisch, S. A., Pierrehumbert, J., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory*, 22(1), 179–228.

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language*, 77(1), 91–106.

Gafos, A. I. (1999). *The articulatory basis of locality in phonology*. New York and London: Garland Publishing, Inc.

- Gafos, A. I., & Kirov, C. (2010). A dynamical model of change in phonological representations: the case of lenition. In I. Chitoran, C. Coupé, E. Marsico & F. Pellegrino (Eds.), *Phonological Systems and Complex Adaptive Systems: Phonology and Complexity*. Berlin/New York: Mouton de Gruyter.
- Galantucci, B., Fowler, C. A., & Goldstein, L. M. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception, & Psychophysics*, 71(5), 1138–1149.
- Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13(3), 361–377.
- Gallagher, G., & Coon, J. (2009). Distinguishing total and partial identity: Evidence from Chol. *Natural Language and Linguistic Theory*, 27, 545–582.
- Gelman, A., & Hill, J. (2007). *Data Analysis using regression and multilevel/hierarchical models*. Cambridge/New York: Cambridge University Press.
- Gordon, P. C., & Meyer, D. E. (1984). Perceptual-motor processing of phonetic features in speech. *Journal of Experimental Psychology: Human Perception and Performance*, 10(2), 153–178.
- Gow Jr., D. W., & Segawa, J. A. (2009). Articulatory mediation of speech perception: A causal analysis of multi-modal imaging data. *Cognition*, 110, 222–236.

- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3), 424–438.
- Greenberg, J. (1950). The patterning of root morphemes in Semitic. *Word*, 5, 162–181.
- Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102(3), 594–621.
- Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96, 280–301.
- Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, 105, 611–633.
- Halle, M. (1971). *The sound pattern of Russian*. The Hague/Paris: Mouton.
- Hickok, G. S. (2008). Eight problems for the Mirror Neuron Theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience*, 21(7), 1229–1243.
- Hickok, G. S., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *TRENDS in Cognitive Sciences*, 4(4), 131–138.

- Hickok, G. S., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8, 393–402.
- Higham, D. J. (2001). An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, 43(3), 525–546.
- Hock, H. S., Schöner, G., & Giese, M. (2003). The dynamical foundations of motion pattern formation: Stability, selective adaptation, and perceptual continuity. *Perception & Psychophysics*, 65(3), 429–457.
- Hoole, P., Zierdt, A., & Geng, C. (2003). *Beyond 2D in articulatory data acquisition and analysis*. Paper presented at the 15th International Congress of Phonetic Sciences (ICPhS XV), Barcelona, Spain.
- Hura, S. L., Lindblom, B., & Diehl, R. L. (1992). On the role of perception in shaping phonological assimilation rules. *Language and Speech*, 35(1/2), 59–72.
- Itô, J., & Mester, R.-A. (1986). The phonology of voicing in Japanese, theoretical consequences for morphological accessibility. *Linguistic Inquiry*, 17(1), 49–73.
- Johnson, K. (2008). *Quantitative methods in linguistics*. Malden, MA: Blackwell Publishing.
- Kamiński, M., Ding, M., Truccolo, W. A., & Bressler, S. L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of signi®cance. *Biological Cybernetics*, 85, 45–157.

- Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language*, 60(2), 286-319.
- Kenstowicz, M. (1994). *Phonology in generative grammar*. Malden, MA: Blackwell Publishing.
- Kerzel, D., & Bekkering, H. (2000). Motor activation from visible speech: Evidence from stimulus response compatibility. *Journal of Experimental Psychology: Human Perception and Performance*, 26(2), 634–647.
- Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, 70(3), 419–4.
- Kirov, C., & Gafos, A. I. (2007). *Dynamic phonetic detail in lexical representations*. Paper presented at the 16th International Congress of Phonetic Sciences (ICPhS XVI), Saarbrücken, Germany.
- Kopecz, K., & Schöner, G. (1995). Saccadic motor planning by integrating visual information and pre-information on neural dynamic fields. *Biological Cybernetics*, 73, 49–60.
- Kornblum, S. (1994). The way irrelevant dimensions are processed depends on what they overlap with: The case of Stroop- and Simon-like stimuli. *Psychological Research*, 56(3), 130–135.
- Krakow, R. A. (1999). Physiological organization of syllables: a review. *Journal of Phonetics*, 27(1), 23–54.

- Ladefoged, P. (1972). *A course in phonetics* (2nd edition ed.): Harcourt Brace Jovanovich.
- Ladefoged, P. (1999). American English. In *Handbook of the International Phonetic Association* (pp. 41–44). Cambridge: Cambridge University Press.
- Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Malden, MA: Blackwell Publishing.
- Levelt, W. J. M. (1992). Accessing words in speech production: Stages, processes and representations. *Cognition*, 42(1–3), 1–22.
- Levelt, W. J. M. (1999). Models of word production. *TRENDS in Cognitive Science*, 3(6), 223–232.
- Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1), 1–38.
- Levelt, W. J. M., & Wheeldon, L. R. (1994). Do speakers have access to a mental syllabary? *Cognition*, 50, 239–269.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431–461.

- Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54*(5), 358–368.
- Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1-36.
- Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*, 384-422.
- Löfqvist, A. (1992). Acoustic and aerodynamic effects of interarticulator timing in voiceless consonants. *Language and Speech, 35*(1, 2), 15–28.
- Lotto, A. J., Hickok, G. S., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *TRENDS in Cognitive Science, 13*, 110–114.
- Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford: Oxford University Press.
- MacEachern, M. R. (1999). *Laryngeal cooccurrence restrictions*. New York: Garland.
- MacKay, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills*. New York: Springer.
- Max, L., & Onghena, P. (1999). Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing

- research. *Journal of Speech, Language, and Hearing Research*, 42(2), 261–270.
- McCarthy, J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry*, 17(2), 207–263.
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1–86.
- Meyer, D. E., & Gordon, P. C. (1985). Speech production: Motor programming of phonetic features. *Journal of Memory and Language*, 24, 3–26.
- Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, 27(2), 338–352.
- Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, 109(1), 168–173.
- Mooshammer, C., Goldstein, L., Nam, H., McClure, S., Saltzman, E. L., & Tiede, M. (2012). Bridging planning and execution: Temporal planning of syllables. *Journal of Phonetics*, 40(3), 374–389.
- Morales, M. (2011). sciplot: Scientific Graphing Functions for Factorial Designs. R package version 1.0-9. <http://CRAN.R-project.org/package=sciplot>.

Müsseler, J. (1995). *Wahrnehmung und Handlungssteuerung. Effekte kompatibler und inkompatibler Reize bei der Initiierung und Ausführung von Reaktionssequenzen*. Aachen: Shaker.

Nielsen, K. Y. (2007). *Implicit phonetic imitation is constrained by phonemic contrast*. Paper presented at the 16th International Congress of Phonetic Sciences (ICPhS XVI), Saarbrücken, Germany.

Ohala, J. J. (1993). Sound change as nature's speech perception experiment. *Speech Communication*, 13, 155–161.

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America*, 99(3), 1718–1725.

Ohala, J. J. (2005). Phonetic explanations for sound patterns. In W. J. Hardcastle & J. M. Beck (Eds.), *A figure of speech. A festschrift for John Laver*. (pp. 23–38). London: Erlbaum.

O'Seaghda, P. G., Dell, G. S., Peterson, R. R., & Juliano, C. (1992). Models of form-related priming in comprehension and production. In R. G. Reilly & N. E. Sharkey (Eds.), *Connectionist approaches to natural language processing* (Vol. 1, pp. 373–408). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies, M. L., Garabieta, I., & Jackson, M. T. T. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America*, 92(6), 3078–3096.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, 32, 693–703.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., . . . Roberts, T. P. L. (2000). Auditory cortex accesses phonological categories: an MEG mismatch study. *Journal of Cognitive Neuroscience*, 12(1038–1055).

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Attention, Perception, & Psychophysics*, 15(2), 285–290.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, 9(2), 129-154.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, 103(20), 7865–7870.

R development core team. (2010). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, <http://R-project.org>.

Rastle, K., Croot, K. P., Harrington, J. M., & Coltheart, M. (2005). Characterizing the motor execution stage of speech production: Consonantal effects on delayed naming latency and onset duration. *Journal of Experimental Psychology: Human Perception and Performance*, 31(5), 1083–1095.

- Rastle, K., & Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2), 307–314.
- Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, 92(1), 81–110.
- Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, 64, 249–284.
- Roelofs, A. (1999). Phonological segments and features as planning units in speech production. *Language and Cognitive Processes*, 14(2), 173–200.
- Roelofs, A. (2000). WEAVER++ and other computational models of lemma retrieval and word-form encoding. In L. R. Wheeldon (Ed.), *Aspects of Language Production* (pp. 71–114). Philadelphia: Psychology Press.
- Rogers, M. A., & Storkel, H. L. (1998). Reprogramming phonologically similar utterances: The role of phonetic features in pre-motor encoding. *Journal of Speech, Language, and Hearing Research*, 41(2), 258–274.
- Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4), 333–382.
- Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, 25, 421-436.

- Sevald, C. A., & Dell, G. S. (1994). The sequential cuing effect in speech production. *Cognition*, 53(2), 91–127.
- Schutte, A. R., Spencer, J. P., & Schöner, G. (2003). Testing the Dynamic Field Theory: Working memory for locations becomes more spatially precise over development. *Child Development*, 74(5), 1393–1417.
- Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial-order mechanism in sentence production. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*. Somerset, NJ: John Wiley and Sons, Inc.
- Shaw, J. A. (2010). *The temporal organization of syllabic structure*. PhD, New York University, New York.
- Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, 66(3), 422–429.
- Schöner, G., Kopecz, K., & Erlhagen, W. (1997). The dynamic neural field theory of motor programming: Arm and eye movements. In P. Morasso & V. Sanguineti (Eds.), *Self-organization, computational maps, and motor control* (1 ed., pp. 271–310). North Holland: Elsevier.
- Smith, C. L. (1995). Prosodic patterns in the coordination of vowel and consonant gestures. In B. Connell & A. Arvaniti (Eds.), *Papers in Laboratory Phonology IV: Phonology and phonetic evidence* (pp. 205–222). Cambridge: Cambridge University Press.

Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. Wathem-Dunn (Ed.), *Models for the Perception of Speech and Visual Form*. Cambridge, MA: MIT Press.

Summerfield, Q., & Haggard, M. (1977). On the distinction of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America*, 62(2), 435–448.

Thelen, E., Schöner, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences*, 24, 1–86.

Tilsen, S. (2007). Vowel-to-vowel coarticulation and dissimilation in response-priming. *UC-Berkeley Phonology Lab Annual Report*, 416–458.

Tobin, S., & Nam, H. (2010). *Asymmetries in Spanish-English gestural drift: Data and model*. Paper presented at the Laboratory Phonology 12, Albuquerque, NM.

Tyler, M. D., Tyler, L., & Burnham, D. K. (2005). The delayed trigger voice key: An improved analogue voice key for psycholinguistic research. *Behavior Research Methods*, 37(1), 139–147.

van Alphen, P. M., & McQueen, J. M. (2006). The effect of Voice Onset Time differences on lexical access in Dutch. *Journal of Experimental Psychology: Human Perception and Performance*, 32, 187–196.

- Viviani, P. (2002). Motor competence in the perception of dynamic events: a tutorial. In W. Prinz & B. Hommel (Eds.), *Common mechanisms in perception and action: Attention and performance XIX* (pp. 406–442). Oxford/New York: Oxford University Press.
- Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology, 24*(5), 672–683.
- Whalen, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics, 18*, 3–35.
- Yaniv, I., Meyer, D. E., Gordon, P. C., Huff, C. A., & Sevald, C. A. (1990). Vowel similarity, connectionist models, and syllable structure in motor programming of speech. *Journal of Memory and Language, 29*(1), 1–26.
- Yuen, I., Brysbaert, M., Davis, M. H., & Rastle, K. (2010). Activation of articulatory information in speech perception. *Proceedings of the National Academy of Sciences (Social Sciences), 107*(2), 592–597.