EXPERIMENTAL AND COMPUTATIONAL INVESTIGATIONS OF F0 CONTROL

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

SeungEun Kim

December 2022

EXPERIMENTAL AND COMPUTATIONAL INVESTIGATIONS OF F0 CONTROL

SeungEun Kim, Ph.D.

Cornell University 2022

This dissertation examines speakers' cognitive control of F0, by proposing and evaluating *target*-control and *register*-control hypotheses. In the *target*-control hypothesis, it is individual pitch targets that speakers control to produce variations in F0, whereas in the *register*-control hypothesis, it is the control of pitch register (in which the pitch targets are defined) that induces F0 variations. These alternative hypotheses are assessed through a production experiment and computational modeling.

The production experiment investigates speakers' (i) pre-planned and (ii) adaptive F0 control. In particular, the experiment examines whether speakers vary F0 parameters (i) according to the initially planned sentence length and (ii) in response to unanticipated changes in the length. For this purpose, a novel experimental paradigm was developed in which the stimuli cueing the parts of the utterance were delayed until after participants initiated an utterance; in this case, participants had to dynamically adapt to the changes in the length and content of the utterance. Analyses of F0 trajectories found strong evidence for both pre-planned and adaptive control. Further analyses were conducted to identify which specific F0 parameter was controlled (*targets* vs. *register*), and the results demonstrated the control of pitch register.

In the modeling study, a gestural model of F0 control was proposed and evaluated with the experimental data. The main feature of this dynamical model is that the normalized targets of F0 gestures (and F0 tract variable) are mapped to actual F0 values through pitch register parameters. The model parameters were optimized to minimize the difference between the empirical F0 contour and the model-generated contour. Several variants of F0 models were compared to examine the *target* vs. *register*-control hypotheses. The results found that the F0 model in which the register parameters varied (with invariant targets) outperformed the model in which the target parameters varied (with constant register), providing further support for the *register*-control hypothesis.

Overall, this dissertation provides evidence that for a given utterance, speakers have a set of invariant cognitive representation of high and low pitch targets, and they control pitch register to realize the abstract representation into different F0 peaks and valleys.

BIOGRAPHICAL SKETCH

SeungEun Kim was born in Daegu, South Korea in 1991. She went to Taegu Foreign Language High School, where she developed her interests in learning and studying different languages. She then moved to Seoul, South Korea and attended Yonsei University. She earned her Bachelor's Degree in English Language and Literature in 2014 and her Master's Degree from the same department in 2017, where she studied psycholinguistics.

In August 2017, she moved to the United States to start her Ph.D. in Linguistics at Cornell University. Her research focuses mainly on phonetics and phonology, with special interests in acoustics, articulation, speech planning, and prosody. In January 2023, she will join the Department of Linguistics at Northwestern University as a Postdoctoral Fellow.

This dissertation is dedicated to my parents.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Sam Tilsen. Sam has taught me how to become a researcher. With his guidance, I learned how to write code, design experiments, analyze data, and present my work in verbal and written forms, all of which have become valuable assets to me. He also supported me emotionally whenever I was skeptical about my work or myself, believing in my potential more than I did myself at times. I am certain that I could not have come this far without him, and I feel so grateful to have had him as my advisor.

I would like to also extend my gratitude to my committee members, Abby Cohn and Draga Zec. Abby always emphasized looking at the "big picture", to see the forest for the trees. Through Abby, I developed skills to connect my work with the literature and to think about its broader implications. Draga provided valuable inputs on the phonological side of my dissertation and my second qualifying paper. Thanks to Draga, I not only built strong foundations in theoretical phonology but also learned how to think critically about them. I also enjoyed teaching Introduction to Phonetics and Phonology with Draga; it was the first linguistics class that I taught at Cornell, and Draga helped me in all aspects of navigating the course.

Outside of my committee, I feel grateful to my American *hal.a.pe.ci*, John Whitman for his support throughout my entire time in graduate school. He was the person I could be candid with about everything, and he always stood by my side. I loved all the conversations that we had together in Korean and English, and it really made me feel at home. Thank you also to Mats Rooth and Marten van Schijndel, who served as chair and committee member, respectively, for my qualifying papers, where I developed my understanding of semantics, computational linguistics, and psycholinguistics.

I feel extremely lucky to have met wonderful people in the Cornell Linguistics department. My first thanks go to my cohort and the phonetics lab members. Thank you to Francesco Burroni and Siree Maspong for caring for me from wherever you were – in Ithaca, Thailand, or Italy. Thank you to Rachel Vogel, who was never hesitant to offer me help in class assignments, read over my writings, and give me advice when I was struggling. A tremendous thank you to Chloe Kwon and Katie Blake, who have been with me in every step of my life (both academically and personally) over the past few years, and who truly care about my happiness more than anything else. I also want to acknowledge Jenny Tindall and Bruce McKee, who took care of all the bureaucracy and technical difficulties that I had to face during graduate school.

I also built true friendships with people outside the Linguistics department at Cornell. Elina Yewon Hur, who will soon become a Ph.D. in Marketing, made my life in Ithaca much more enjoyable. The conversations we had over numerous coffees, dinners, and wines, and the pilates sessions we had together were precious breaks that gave me moments to breathe and feel present in the midst of hectic Ph.D. life. Hankyul Kim is another person to whom I should express my gratitude. He cared for me like an older sibling; whenever I asked for something like an annoying little sister, he was always there to help me (although it was often accompanied by a little bit of complaint). I cannot imagine surviving in Ithaca without Elina and Hankyul. During my Ph.D. program, I also had so many great students in both my Linguistics and Korean language courses. They were more than just students to me; they made me laugh and brightened my stressful days. Special thanks to Jamie Wang and Cole Horvath, who have been my best student friends from their freshmen years to even after graduation. I would also like to thank my Korean supervisor, Meejeong Song, and Cornell East Asia program for the fellowship.

There were also many friends and mentors who helped me throughout my undergraduate and Master's programs in Korea. I would first like to mention my dear professors in the Department of English Language and Literature at Yonsei University. Professors Seok-Chae Rhee, Seung-Hee Lee, and Eun-Kyung Lee introduced me to the field of Linguistics and supported me in every way when I decided to go to the United States for my graduate studies. My friends whom I met in my freshman year at Yonsei – Kyeongryang Kim, Ayeong Kim, YoonYoung Park, Hyojung Paik, Minhye Eom, Yeonjung Hong – are my greatest life assets, which I would never give up for anything. I also want to mention Heeyeon Kim, Euimi Park, Min Kyung Yang, and Sehyung Lee for their support and our ongoing friendship. All these people texted and called me day and night to make sure that I was doing okay, and that I did not feel lonely in a faraway country. I also deeply appreciate Chunja Koh, who was once my high school teacher but is now my best life mentor.

Finally, words cannot express my gratitude for my family. My parents, Ki Won Kim and Taesuk Jun, have been great life-long supporters of their daughter. I grew into the person I am today thanks to these two incredible people. They have believed in me and supported every life choice I have made. They have always told me that they will be there for me whenever things go wrong, which made me strong and move forward with my life. I also feel grateful to my grandma, who has been always so proud of me, as well as my grandpa, who passed away a few years ago but who I believe is still watching over me from a better place. Thank you to my little brother and sister-in-law, Jong Heon Kim and So Yeon Oh, for putting up with this picky sister.

All my academic achievements and life journey until now have been possible with the help of so many wonderful people around me. I feel extremely grateful that I am surrounded by an incredible support system, including those mentioned here and those who were not.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
GENERAL INTRODUCTION

Imagine a situation where A and B are talking about their mutual friend, *Jamie*. Speaker A says, "*Jamie is from Delaware.*" Speaker B is skeptical about this information (as it is different from what he remembers) and asks, *"Is she?"* This leads to Speaker A correcting her statement, where she says *"Sorry, Jamie is from Maryland.*"

One can easily predict that the intonation of two utterances from the same Speaker A – i.e. *"Jamie is from Delaware" vs. "Jamie is from Maryland"* – would significantly differ. The most prominent difference would be that the maximum fundamental frequency (F0) value of *Maryland* (in the correcting sentence) would be higher than that of *Delaware* (in the original sentence), although they are both tri-syllabic words with initial stress, and they occupy the same syntactic position in the sentence.

What does Speaker A do to produce different F0 patterns in these contexts? More fundamentally, how is this difference manifested in the Speaker A's cognitive control system of F0? This dissertation is motivated by the insight that there are two alternative ways in which the control of F0 can be conceptualized. One is that the speaker controls individual *pitch targets*; in the example above, the speaker has a higher pitch target for *Maryland* than for *Delaware* – i.e. the speaker aims to achieve a higher F0 peak in the first syllable of *Maryland*. I will refer to this idea as *target*-control in this dissertation. The other is that the speaker controls *pitch register*, the F0 space in which the pitch targets are defined; in the example, the speaker either broadens the F0 space around *Maryland* or shifts it up – i.e. the speaker expands/shifts the space (not the targets), and

the targets are realized within that space. I will refer to this possibility as *register*-control henceforth. Note that "F0" and "pitch" are used interchangeably throughout this dissertation.

The current study investigates how speakers control F0, specifically, whether they mainly control *pitch targets* or *pitch register* to produce variations in F0. This is an important question to ask, as it can suggest what ingredients we need to account for F0 contours – i.e. how F0 contours can be decomposed and what are their building blocks. It would also inform us about the cognitive mechanisms that drive various empirical F0 phenomena that have been attested in the literature – for instance, downstep or declination. Yet, despite its significance, the question of how speakers control F0 has been largely ignored in the field. Although pitch targets and register are considered as important notions in understanding F0 contours, studies have been vague about what it is that speakers *control* to produce various F0 peaks and valleys. It is thus important to explicitly delineate the possible hypotheses of pitch control – i.e. *target* vs. *register* control – and further examine them empirically.

In this dissertation, the two hypotheses of F0 control were evaluated first through a production experiment and then through a modeling study. In the experiment, I examined how speakers vary F0 parameters – specifically, peaks, valleys, and their ranges – according to the initially planned sentence length as well as how they respond to changes in the length that are made after utterance initiation. To identify which F0 parameter (i.e. *target* vs. *register*) speakers mainly control to produce such variations, analyses of the variance and correlation of F0 measures as well as the comparison of condition-prediction models were further conducted. In the modeling study, the gestural

model of F0 control which was developed in the framework of Articulatory Phonology was introduced and tested on the empirical data collected from the experiment. Crucially, different variants of models which exemplified the *target* and *register*-control hypotheses were constructed, and their performances were compared.

Overall, this dissertation provides evidence that speakers control pitch register to produce variations in F0. The experiment results found that participants make a pre-utterance F0 plan which takes the initial sentence length into account, and moreover, they can adjust that plan online in response to the changes in the length. Furthermore, analyses of variance and correlation of F0 measures provided evidence for the *register*-control hypothesis. This was further confirmed in the modeling, where the F0 model that mainly varied pitch register outperformed the model that varied pitch targets, lending support to the *register*-control hypothesis. Below, I introduce the key concepts of this dissertation – *pitch targets* and *pitch register* – and elaborate the main hypotheses of speakers' F0 control – *target*-control vs. *register*-control. I also briefly introduce the goals and methods of the production experiment and computational modeling and present the organization of this dissertation.

1.1 Pitch target

In this dissertation, I use the phrase "pitch target" in a couple different ways. For one, in a general sense, pitch target refers to the idea that speakers have some cognitive representation of what they want their F0 to be while they are speaking. It is a theory- and model-neutral concept in that it makes minimal assumptions about the nature of the

control system. The main idea is that the speakers have an F0 goal or target, and the vocal control system has a parameter that expresses this value. The other usage of "pitch target" in this dissertation is more specific in the sense that it refers to the values of parameters in a dynamical F0 model, which is presented later on. The parameter value is defined in abstract/normalized coordinates in the interval from 0 to 1, and it reflects speakers' cognitive representations of where F0 should be within this range. Figure 1.1 provides a schematic representation of pitch targets and register used in this dissertation; the red and blue dots indicate the high and low pitch targets, respectively.

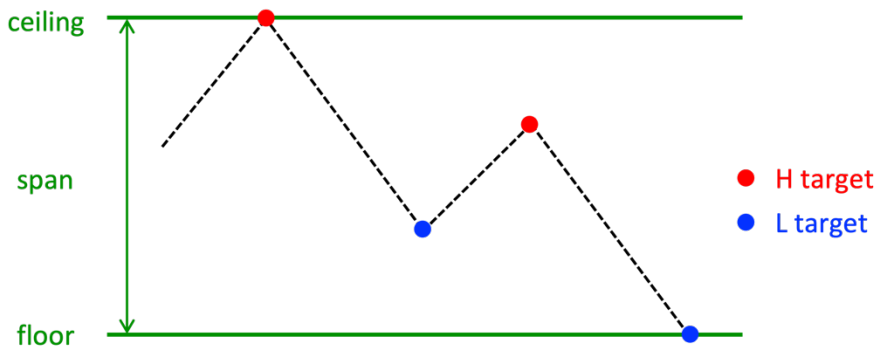

Figure 1.1. A schematic representation of pitch targets and pitch register used in this dissertation. The red and blue dots mark high and low pitch targets, respectively. The green lines show pitch register ceiling, floor, and span; see Section 1.2.

In the literature, the notion of pitch target has been conceptualized in various ways. For instance, in the Autosegmental-Metrical intonation model, it was considered that speakers aim for distinctive pitch levels (points), and these targets are realized as peaks and valleys in F0 contours (e.g. Pierrehumbert, 1980). The PENTA model proposed by Xu (2005) also assumed pitch targets to be specific levels or rises/falls, which however are underlying targets that do not necessarily map to the peaks and valleys in the surface F0 contours. On the other hand, some computational F0 models considered pitch targets

as movements, which then require parameters that specify target value as well as the shape of the movement. Specific details on how various F0 models proposed in the literature define "pitch targets" and how these targets are implemented will be discussed in Section 2.1.

1.2 Pitch register

I also use "pitch register" in both a general and specific sense in this dissertation. In the general sense, it refers to a range of F0 values that speakers can produce and utilize at a given time in an utterance. This notion allows for potential changes in the range – i.e. the range can be expanded, compressed, or shifted. Note that the F0 range discussed in this dissertation is not a "physiological" range, but the range that speakers use in their communication[1]. Since pitch register embodies the notion of range or space, it is defined by a combination of at least two of the following three parameters: ceiling (topline), floor (bottomline), and span (range). The more specific sense that I adopt is associated with a control architecture in the dynamical F0 model. In the model, pitch register is a set of parameter values that map normalized (or abstract) F0 coordinates to actual F0 values. The green horizontal lines in Figure 1.1 indicate the register ceiling and floor, and its range represents the span.

[1] In some studies, the notion/term of "pitch register" has been distinguished from "pitch range". For instance, Ladd (1990) argued that pitch range is considered as constant for a given speaker on a given occasion, while pitch register is a subset of the pitch range and refers to a band of F0 values relative to which the tonal events are scaled. While pitch register can vary at a phrase or sentence boundary or can be used to mark local prominence, pitch range is varied by paralinguistic factors such as overall interest or arousal. Under this perspective, pitch register does not exceed pitch range.

The concept of pitch register was introduced in the F0 literature under different names. One term that was proposed by Ladd (1992) is "tonal space", in resemblance to "vowel space". Both tonal space and vowel space delimit the region where F0 and vowels can be produced, and individual tones and vowels are understood given the positions that they occupy within the space. Moreover, tonal space and vowel space can expand or compress depending on the context. Besides tonal space, "tonal level frame" (Clements, 1979), "transform space" (Pierrehumbert & Beckman, 1988), or "grid" (Gårding, 1983) have also been proposed in the literature (Ladd, 2008). While all these terms represent the notion of space/range as pitch register in this dissertation, some computational F0 models lack a full specification of range, but instead merely specify a rough position within the range (for example, in the form of a line) that F0 targets are superimposed on. The details of these models will be introduced in Section 2.1.

1.3 F0 control hypotheses

A crucial insight of this dissertation is that for any utterance with an F0 trajectory, there is a basic ambiguity on what the speaker has controlled to produce that trajectory. This ambiguity gives rise to two possible hypotheses of F0 control, the validity of which is examined and compared in this dissertation: these hypotheses are (i) *target*-control and (ii) *register*-control hypotheses[2]. I illustrate each of these hypotheses below with reference to Figure 1.2.

[2] In this dissertation, the two hypotheses are evaluated on the intonation language (i.e. English). Whether the results of this study can be extended to tone languages is left for future works (see Chapter 5).

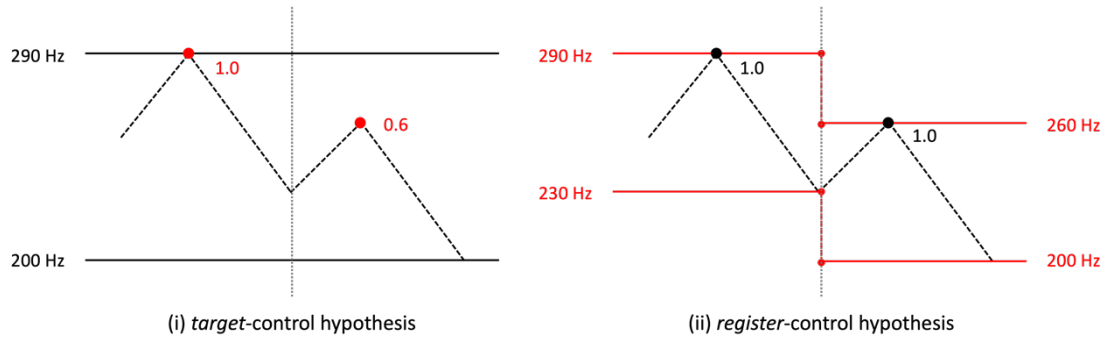

Figure 1.2. Schematic illustrations of pitch control hypotheses that are investigated in this dissertation. In the figures, the black dashed line represents a sample schematic F0 contour, the solid horizontal lines show pitch register ceiling and floor, and the dots at the F0 peak indicate high pitch targets. It is assumed that there is a prosodic boundary (prosodic word or phrase) between the two peaks, which is indicated as a vertical dotted line. The left side of the figure shows the mechanisms under the (i) target-control hypothesis, and the right side shows the (ii) register-control hypothesis. The main variables that lead to variations in F0 under each hypothesis are marked in red. An arbitrary parameter value is presented for the high pitch target (1.0, 0.6) and floor (200, 230 Hz) and ceiling (260, 290 Hz) to promote understanding of the control parameters of each hypothesis.

Under the (i) *target*-control hypothesis, variation in the values of the F0 peaks and valleys within an utterance is directly manipulated by speakers. Figure 1.2 presents how two alternative hypotheses account for the realization of two distinct F0 peaks. In the left side of Figure 1.2 – (i) *target*-control hypothesis, when speakers are producing an F0 contour with two different peaks, they would have two distinct high pitch targets in mind: one is a higher target, for example at the top of the F0 range available for a given utterance, and the other is lower than the previous one, for example at around 60% of the range. If we refer to this range on a scale from 0 to 1 (this allows us to abstract away from the specific differences between speakers and contexts), the first high target would have a value of 1.0, and the second target as 0.6. The reader, however, should note that the specification of the targets in the register-normalized units (from 0 to 1) is not

essential to the idea of the *target*-control, although I use it here and elsewhere to be consistent with the model developed in Chapter 4. Although the figure only shows values for high pitch targets, speakers would likewise have distinct low pitch targets to produce different F0 valleys. (cf. However, I do not assume any symmetry in the control of high and low targets.) In this hypothesis, pitch register remains constant throughout the utterance, which is shown as the black solid lines, and their values are identical across prosodic units (200-290 Hz).

Under the (ii) *register*-control hypothesis, variation in the F0 values of the peaks and valleys within an utterance is considered to arise from changes in register. In other words, speakers adjust tonal space, while F0 targets – expressed relative to register – remain constant. The observed F0 values of peaks and valleys can thus be understood as the indirect consequences of pitch register variation. See the right side of Figure 1.2. Pitch register is shifted downwards at the prosodic boundary (from 230-290 Hz to 200-260Hz), and this results in distinct F0 peak values. It is important that although the surface F0 values of the two peaks are different (290 vs. 260 Hz), F0 targets stay the same (both are 1.0). This means that speakers have the same high pitch target across prosodic units (which in this figure is at the top of the range), but the control of register results in different surface F0 peaks.

There are a couple of crucial points to make regarding the illustrations of the hypotheses. The first is that the surface F0 values of the peaks and valleys we observe in utterances cannot be assumed to directly represent targets of speaker control. There are three reasons for this. First, and perhaps most importantly, we do not have direct access to the F0 control system that speakers use when producing speech, and thus we

cannot simply assume that targets are defined in a familiar, physically measurable unit of Hz or some transformations thereof (such as ERBs, semitones, etc.). Second, as a consequence of rejecting this assumption, it becomes logically possible that variation in F0 values arises from adjusting the register, understood in the technical sense as a mapping from a control space to a range of F0 values. If one allows for this form of control, it is possible for F0 variations to arise even when cognitive targets remain constant, if the register itself changes. This is in fact the essence of the (ii) *register*-control hypothesis. Third, it is important to note that we cannot always assume that speakers achieve their targets, whether these are governed more directly via target control or more indirectly through register control. Thus, the F0 values that we observe in speech might fail to reach the intended values of the control system, in a form of target undershoot. Namely, the fact that F0 trajectories show peaks and valleys does not entail that the values of those peaks and valleys are themselves the targets – i.e. peaks and valleys can arise whenever a control system prematurely switches from one target or register to another. However, it does seem reasonable to allow that in speech with relatively moderate tempo, target undershoot of this sort is less likely, in which case, F0 values of peaks and valleys can be more safely assumed to reflect the underlying intentions of the control system.

The other important point is that the two hypotheses that are investigated in this dissertation are not necessarily exclusive. Variation in F0 might arise from both target *and* register control – i.e. the combination of the two hypotheses. It is, however, not possible to fully resolve whether both forms of control are used, nor the extent to which one or the other predominates. Thus, as a first step of investigating speakers' F0 control

mechanism, it makes sense to restrict to the hypotheses which allow variations in one set of control parameters (i.e. *target* vs. *register*).

The attempt to resolve between the two hypotheses of F0 control with empirical data is quite challenging and to my knowledge has not been tried before. Although the concepts of pitch targets and register have been widely discussed in the F0 literature, no studies have indeed tried to examine what is controlled by speakers to produce variations in F0. Many of the arguments I will bring to bear on this are necessarily indirect and ultimately may not be conclusive. Nonetheless, it is important to recognize that there is an inherent ambiguity in what speakers are controlling, and this dissertation contributes to our understanding of F0 production by developing experimental and computational methods that may be used to resolve the ambiguity.

1.4 Production experiment

To assess our main hypotheses of pitch control, a sentence production experiment was conducted. The experiment used a novel paradigm in which the length and content of the utterance were adjusted *after* speakers initiated the utterance. I refer to this as an *adaptive* control task, because under some conditions, speakers must adaptively adjust to the changes in the length and content of the utterance that they are required to produce, with those changes being cued after they have started production.

The point of introducing this manipulation is that it perturbs the control system, so that it may inform us about what it is that speakers are controlling to produce F0 variations. The paradigm can be understood as an instance of the more general class of

perturbation studies, such as feedback perturbation, except that instead of perturbing auditory feedback during speech, this experiment perturbs the utterance plan itself; in this sense, some general introductions on studies that involved F0 perturbations via auditory feedback will be presented in Section 2.2.4.

Also, the current experiment design is similar to the one employed in Whalen (1990), who examined coarticulation in the context where speakers are asked to produce a sequence of syllables with a segment missing initially. Participants, for instance, read a nonsense disyllabic sequence (ɑ**bí**, ɑ**bú**, ɑ**pí**, ɑ**pú**), in which either the consonant (b/p) or vowel (i/u) was missing at the beginning of the trial. The missing segment was initially presented as a blank (e.g. A_I or AB_ in the first example), but it immediately appeared when participants initiated a response. Participants were instructed to incorporate the delayed segment into the ongoing production as rapidly and smoothly as possible. The results did not find the anticipatory coarticulation associated with the missing segment (e.g. the lengthening of /ɑ/ according to the following /b/ or /p/), but still found the perseverative coarticulation (i.e. F0 of /i/ and /u/ affected by the preceding /b/ or /p/). A similar study in which segmental planning was perturbed by delaying parts of the necessary stimulus is Tilsen (2014). Their findings suggest that the participants in the current study may be able to control F0 according to the perturbations, yet the perturbations in this study are more fundamental in that they alter the length and content of the utterance.

One might object that the perturbation of the utterance plan is too unnatural or too different from the conditions of normal conversational speech. In some ways, this is the point: by imposing an unusual perturbation on the system, we can draw inferences that

might otherwise be difficult or impossible to obtain. However, I would also note that the unanticipated changes in utterance length/content may in fact be quite common in spontaneous speech; speakers may decide to change their utterance plan after they have begun the utterance for a variety of reasons – for instance, some new stimuli appear in the environment, or their internal monitoring of the message detects an error or omission and thereby induces a revision.

In addition to testing speakers' ability of *adaptive* F0 control, the experiment also examined the effect of sentence length on the *initial/pre-planned* F0 control. The previous literature has been inconclusive on whether speakers make a pre-utterance plan that considers the length of the sentence that they are going to produce. In particular, studies have examined whether speakers adjust the initial F0 as a function of utterance length, by testing a hypothesis that they raise their utterance-initial F0 peak in longer sentences. The hypothesis was tested in a variety of languages, but the results differed within and across languages; Section 2.2.3 presents the summary of findings of these studies. This dissertation aims to reexamine this question from the lens of what it is that speakers are controlling to produce variations in F0.

In the experiment, participants produced sentences that differ in length. In particular, the length of the subject phrase was varied in that it was composed of one, two, or three conjoined noun phrases (NPs). An example for the three NP sentences is *"Nine green rhinos and eight red weasels and eight blue llamas live in the zoo*". In half of the sentences that had two or three NPs, the non-initial NPs (*"eight red weasels and eight blue llamas"*) were presented not at the beginning of the trial, but after participants initiated production; I refer to this manipulation as *delayed-stimuli*. An example of a

delayed-stimuli trial with three subject NPs is shown in Figure 1.3-(A); Figure 1.3-(B) shows an example of a no-delayed stimuli trial, where all visual stimuli were presented at the beginning of the trial. A total of five experimental conditions – i.e. 1NP, 2NP with/without delayed-stimuli, 3NP with/without delayed-stimuli – were tested, and 13 native speakers of English participated in the experiment.

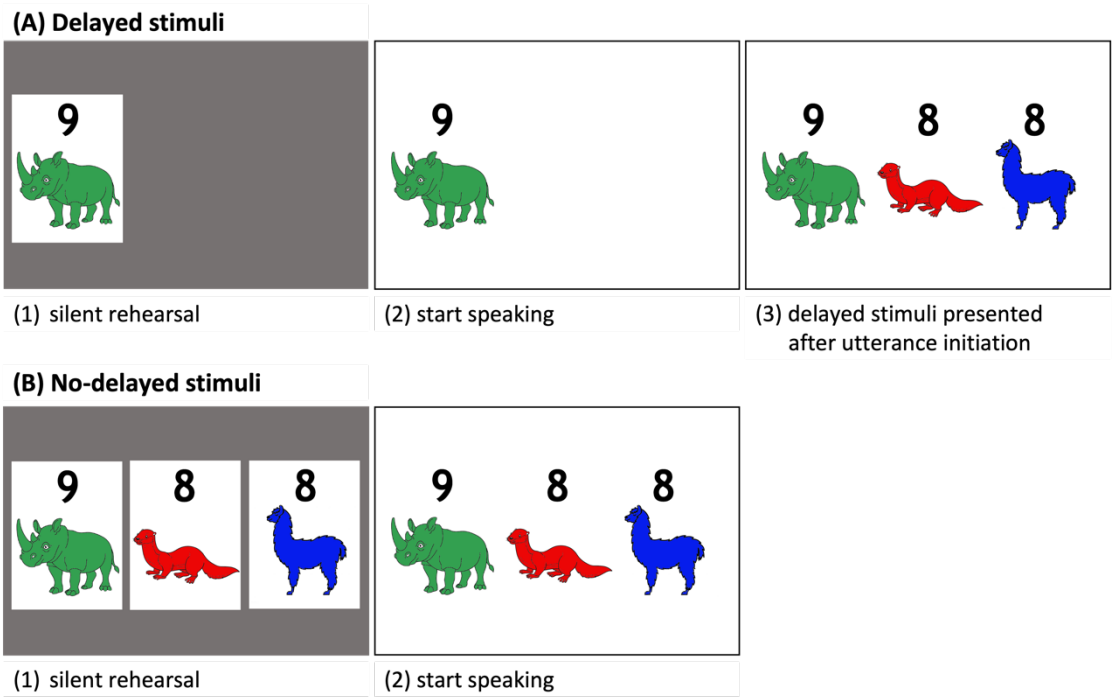

Figure 1.3. Presentation of a single trial with three subject NPs (A) with and (B) without delayed-stimuli. All NPs were presented with visual cues. (1): The initial stimuli were presented with a grey background, and participants were instructed to silently rehearse the sentence during this phase. (2): After some periods of time, the grey background automatically changed to white, and participants could start speaking. (A)-(3): In trials with delayed-stimuli, the visual stimuli that cued non-initial phrases appeared as soon as the utterance initiation was detected; participants were instructed to incorporate the delayed phrases into their ongoing utterance. Further details on the experiment methods are presented in Section 3.2.

The first set of the analyses aimed to examine how participants vary F0 parameters according to the experiment manipulations – i.e. the initial sentence length and changes

in the length. A majority of participants produced F0 trajectories where there is an F0 valley, an F0 peak, and another F0 valley at each NP. The key measurements for this analysis were therefore the F0 values of the peaks, the valleys preceding/following the peaks, and their ranges. If participants pre-plan F0 according to the initial sentence length, F0 parameters would significantly differ among trials in which one, two, or three NPs are initially presented. In addition, if participants adapt to the changes in the utterance length, they would control F0 of the trials with vs. without delayed-stimuli differently after they encounter the delayed stimuli.

The second set of the analyses aimed to examine which parameter (i.e. *target* vs. *register*) speakers specifically control to produce variations in F0. The challenge for this analysis is that we cannot directly measure pitch targets or pitch register from F0 trajectories. Therefore, the analyses focused on the variables that can be interpreted as indirect reflections of underlying target or register parameters. Under the *target*-control hypothesis, F0 values measured at peaks and valleys can be interpreted as F0 targets. In contrast, under the *register*-control hypothesis, the values of peaks and valleys can be interpreted as approximations of (or at least, correlates of) the ceiling and floor of the register, with their difference being a correlate of the span. Since F0 peaks and valleys can either be interpreted as high/low F0 targets or ceiling/floor, I consider the difference between the peaks and valleys (i.e. ranges/span) as a better reflection of pitch register.

To determine which hypothesis of F0 control better accounts for the variations observed in the data, I first examined the variance of F0 measures, comparing the sum of the variances of F0 peaks and valleys and the variance of ranges. Furthermore, I examined the correlation between F0 peaks and valleys within each NP. Lastly, I

compared three regression models that had either F0 peaks, valleys, or ranges as an independent variable and examined which model best predicts the delay vs. no-delay experimental condition. It is important to emphasize again that the ability to assess our hypotheses through empirical data depends on the accuracy of F0 measures and the assumption that they faithfully reflect the underlying control parameters. Therefore, inferences that are drawn regarding the hypotheses are necessarily conditional on various assumptions.

Overall, the results from the experiment showed evidence for the speakers' pre-planned and adaptive F0 control. Specifically, participants varied F0 parameters according to the initially planned sentence length and further adjusted the parameters to adapt to the changes in the length. The analyses of F0 control mechanism provided support for the *register*-control hypothesis.

1.5 Computational modeling

To further examine our main hypotheses of F0 control, a computational, mathematical model that is based on the framework of Articulatory Phonology (AP) and Task Dynamics (TD) was developed. Previous descriptions of F0 control in AP framework include Gao (2008), Mücke et al. (2012), Katsika et al. (2014), and Yi (2017), the details of which are introduced in Section 4.1. In the AP/TD framework, F0 is viewed as a tract variable, directly analogous to other tract variables (except that the F0 is not given an explicit articulatory correlate) such as lip aperture (LA) or tongue tip and tongue body constriction location and degree (TTCL, TTCD, TBCL, TBCD), and

the fundamental units of F0 are F0 gestures. Changes in the F0 tract variable occur when F0 gestures become active, and these active gestures change the equilibrium of the F0 tract variable.

A gestural model of F0 control, proposed for the first time in this dissertation, has two main components. One is the target parameter of F0 gestures, which is expressed in the normalized coordinates in the range from 0 to 1. The targets of F0 gestures along with neutral attractors determine the dynamic target of the F0 tract variable and result in F0 movements. The other main component of the model is pitch register parameters – specifically, the floor and span parameters. These parameters map the normalized F0 coordinates into actual F0 values in the unit of Hz; in particular, the normalized F0 coordinates are multiplied by the value of the span parameter and then added to the value of the floor parameter. The details on the mechanisms of the model and other parameters are introduced in Section 4.2.

Relating this model characterization with our hypotheses of pitch control, the targets of F0 gestures were mainly varied to implement the *target*-control hypothesis, whereas the register parameters were varied to implement the *register*-control hypothesis. Specifically, one high (H) and one low (L) gesture were posited in each NP, since the general form of F0 trajectories had an F0 valley – peak – valley in each NP. The *register*-control hypothesis was implemented by setting the targets of H and L gestures identical across NPs (i.e. H/L targets of the NPs have the same value), but allowing the register parameters to vary at a phrasal boundary, as in Figure 1.2-(ii). The values of the gestural targets stay the same across phrases, and the varying register drives F0 variations. On the other hand, the *target*-control hypothesis was implemented

by directly varying the H and L gestural target values (i.e. H/L targets of each NP have distinct values), but setting the register parameters constant, as in the schematic representation of Figure 1.2-(i).

These F0 models were fit to the empirical F0 contours collected from the experiment, and the model fits were compared. The model parameters, including gestural targets and register floor/span, were optimized to minimize the root mean squared differences between the input F0 contour and the model-generated contour. The performance of the models was then compared with the root mean squared differences.

The modeling study overall found support for the *register*-control hypothesis. When the performances of the F0 models, which exemplified the *target*-control and *register*-control hypotheses, were compared, the model that allowed variation in the tonal space (with invariant target) outperformed (i.e. had smaller root mean squared differences) the model in which the gestural targets varied across phrases (with constant register).

1.6 Overview of dissertation

The rest of the dissertation is organized as follows. Chapter 2 presents the relevant literature of the current study. Specifically, Section 2.1 introduces the conceptual and computational models of F0 control proposed in the literature. In doing so, I discuss whether each model exemplifies the (i) *target*-control or (ii) *register*-control hypothesis. Section 2.2 presents empirical F0 phenomena – i.e. downstep, declination, initial F0 control with respect to sentence length, and adaptive control in feedback perturbations

– and discusses each phenomenon with respect to our hypotheses of pitch control.

The production experiment and computational modeling are presented in Chapter 3 and Chapter 4, respectively. Chapter 3 starts with an introduction and presents hypotheses and predictions (Section 3.1). Section 3.2 details experiment design and analysis methods, and Sections 3.3 and 3.4 present the results and discuss them. In Chapter 4, Section 4.1 gives an introduction (with an introduction on AP and F0), and Section 4.2 presents the basic mechanisms and implementation details of the gestural model and illustrates how the model is tested. Section 4.3 describes experiment and analysis methods, and Section 4.4 present results, which are further discussed in Section 4.5.

Lastly, Chapter 5 concludes the dissertation, summarizing the main findings of the current study along with some contributions and future directions.

CHAPTER 2
BACKGROUND

In the previous chapter, I introduced the main question of this dissertation: how do speakers control F0? I also presented two logically possible hypotheses on how F0 might be controlled, which are the (i) *target*-control hypothesis and (ii) *register*-control hypothesis.

In this chapter, I first review five different conceptual and/or computational models of F0 production (or generation), which are the (i) Autosegmental-Metrical intonation model (e.g. Liberman & Pierrehumbert, 1984; Pierrehumbert, 1980, 1981; Pierrehumbert & Beckman, 1988), (ii) grid model (Gårding, 1983), (iii) soft-template model (Kochanski & Shih, 2003), (iv) command-response model (Fujisaki, 1983), and (v) PENTA model (Xu, 2005). In the course of elaborating the details of each model, I discuss whether the model exemplifies the *target*-control or *register*-control hypothesis (or possibly both). Note that the control mechanism of these models is identified based on the specific notions of pitch targets and register that are adopted in this dissertation (Chapter 1); in particular, pitch targets are understood as the specific *levels* that speakers want their F0 to be, and register is the F0 *space* within which the targets are defined.

I also introduce empirical F0 phenomena which are relevant to the discussion of our main question. These phenomena are (i) downstep, (ii) declination, (iii) effects of sentence length on the pre-planned, initial F0 control and (iv) adaptive F0 control with respect to F0 perturbations. In each case, I consider how the empirical pattern can be analyzed under different hypotheses of pitch control.

2.1 F0 models

2.1.1 Autosegmental-Metrical intonational model

The Autosegmental-Metrical (AM) model of intonational phonology (e.g. Ladd, 2008; Liberman & Pierrehumbert, 1984; Pierrehumbert, 1980, 1981; Pierrehumbert & Beckman, 1988) is an example that is consistent with the *target*-control hypothesis. In the AM framework, the abstract, phonological primitives of intonation are high (H) and low (L) tones. These tones are associated with a stressed syllable or a phrasal boundary, each of which is referred to as pitch accents and edge tones (phrase accents or boundary tones).

The H and L tones are phonetically translated as tonal targets, which are the turning points in the surface F0 contours such as F0 peaks and valleys. The rest of the values in the F0 contour are then derived by an interpolation between the tonal targets. The phonetic implementation of the AM model illustrated in Pierrehumbert (1981) specified pitch targets as locations within the given pitch range. In particular, target values were expressed as a fraction of distance from the bottomline (the bottom of the speaker's F0 range) to the topline (the top of the pitch range) in normalized values from 0 to 1, where 0 refers to the baseline and 1 to the topline. The interpolation between the tonal targets was modeled with a quadratic function, which generated a sagging or monotonous form of transitions.

Liberman and Pierrehumbert (1984) proposed phonetic implementation rules to model variations of F0 peaks (especially on their lowering) in the empirical trajectories. They posited parameters such as reference level, downstep constant, and baseline to model F0 contours. Here, the surface F0 values of tonal targets were first transformed

by subtracting the reference line; in their data, F0 peaks exhibited an exponential decay, which headed towards an asymptote, and this asymptote was considered as the reference line for each phrase. This means that the surface measures of F0 peaks were always understood as some distance above the reference level. The reference level is, however, different from the baseline, which is invariant for a given speaker; the reference level is always located above the baseline, and the former varies with changes in pitch range, but not the latter. The downstep parameter was applied to the transformed tonal target (i.e. surface F0 - reference level) and specified how much the current target decreased from the previous target (i.e. phonetic scaling). The model also had a parameter that affects the final accent, modeling the extra lowering observed at the final peak.

The fundamental component that generates variations in F0 under AM framework is the H and L tones (e.g. the compositions of tones, their locations within an utterance, alignment) and the phonetic realization rules (e.g. scaling), but the model also acknowledges the effect of pitch range in F0 production. For instance, Liberman and Pierrehumbert (1984) mentioned that speakers may speak up to be heard at a distance or to make their voice sound cute, which leads to an expansion of pitch range and thereby resulting in an overall scaling of pitch targets; in particular, this was done through changes in the reference level in their model. It is, however, important that the AM model considers pitch range variation to occur only when signaling paralinguistic information. This differs from the register control discussed in this dissertation, which is the core underlying mechanism that drives all sorts of F0 variations. It could thus be summarized that the AM model mainly focuses on the control of pitch targets, yet it allows for the register control, specifically for paralinguistic purposes.

2.1.2 Grid model

Gårding's grid model (1983) can be understood as an instantiation of the *register*-control hypothesis. Gårding (1983) proposed seven rules to generate a pitch contour. The first rule is to draw a tonal grid, which is defined as "the global frame for the sentence intonation within which the local pitch movements can develop". The tonal grid is composed of a ceiling and a floor and specifies the space that can be utilized by speakers in their normal pitch range. It can also expand or compress to signal semantic and pragmatic information (e.g. focus) or vary with sentence type (e.g. statements vs. questions). Within the grid, there is another set of lines which indicate the bounds for accents within a phrase. Thus, in the grid model, two sets of F0 spaces are defined: the first exterior set is used to mark semantic/pragmatic contexts, while the second interior set is used to specify the targets of the accents.

The next rules in the pitch generation algorithm insert highs and lows onto the exterior/interior grid lines to mark word or phrase accents and mark phrase or sentence boundaries; the highs and lows could be understood as the high and low pitch targets in this dissertation. Since the grid lines demarcate where these highs and lows could be located (i.e. the space is defined first and that determines the placement of high and low targets), this is an example of the *register*-control model. In other words, depending on the location of the grid lines or their width, the realization of the abstract highs and lows would differ. The rest of the rules adjust these highs and lows according to the context and ultimately combine all given information to generate an F0 contour.

In sum, the grid model exemplifies the *register*-control hypothesis, as it is mainly the tonal grids that result in the variations in F0, and high and low targets are inserted

onto the grids. The weakness of this model, however, is that it is a conceptual model, which has not been tested quantitatively on a large scale of empirical data.

2.1.3 Soft-template model

The rest of the models introduced in this section – soft-template model (2.1.3), command-response model (2.1.4), and PENTA model (2.1.5) – have both components of pitch target and register control, yet they differ from the target (as pitch *levels*) and register (as pitch *space*) discussed in this dissertation.

The soft-template model proposed by Kochanski and Shih (2003) was mainly developed to improve intonation synthesis in the text-to-speech (TTS) system. The key component of the soft-template model is the mark-up tags, which specify linguistic/prosodic information, and at the same time, generate an F0 contour via a set of mathematically defined parameters that are associated with the tags. To generate a pitch contour, *phrase curve* is first defined. This specifies the location of an F0 contour within the speaker's F0 range as well as the overall shape and direction of the contour; it is defined mainly with <step> and <slope> tags. Next, tones and accents are defined via <stress> tags. Each stress tag specifies the shape and height of the tones and accents, yet its surface manifestation is determined through an interaction with phrase curve and the neighboring tones and accents.

Relating these components of the models with our hypotheses of pitch control, the stress tag can be understood as an exemplification of the pitch target control, as it specifies the targets of tones and accents. However, it should be noted that the stress

tags in this model not only specify the target values but also the shapes of the tones and accents. On the other hand, pitch register control is similar to defining the phrase curve, which specifies where in the range a pitch contour is located as well as its rough shape. This, however, differs from the pitch register of this dissertation, as my use of register control refers to the control of F0 *space* (which is defined with a ceiling, floor, and span), not simply the height of the contour; in other words, the phrase curve of the soft-template model is one-dimensional, whereas the pitch register of this dissertation is the two-dimensional concept.

The soft-template model is a quantitative model, which was evaluated on the intonation of Mandarin Chinese. In particular, Kochanski et al. (2003) modeled F0 trajectories of the Mandarin corpus, where a male speaker read paragraphs from the news articles that included a wide variety of tonal environments. The model showed good performance, as they could fit the F0 contours of the corpus data with just a 13 Hz root mean squared errors (RMSE) and explain 87% of the variance of the data. Yuan (2004) modeled sets of statements and questions in Mandarin Chinese with the soft-template model to investigate intonational differences between declaratives and interrogatives. The model provided good fits of the data, as the RMSE of the best example was 9.4 Hz. They further showed that questions overall have a higher phrase curve than statements and that the sentence-final syllables of questions are more likely to retain their intended pitch shape rather than being affected by the adjacent tones.

2.1.4 Command-response model

The command-response model developed by Fujisaki (Fujisaki, 1983, 2003;

Fujisaki & Hirose, 1984) also exemplifies pitch target and register control, but in a way different from the control mechanism discussed in this dissertation.

The basic assumption of this model is that the surface F0 contour is the response of the phonatory system to two types of suprasegmental commands – i.e. the phrase command and the accent command. The phrase command defines an overall F0 contour shape of a phrase, which is characterized by a slow phrase-initial rise and a gradual, asymptotic decline to a baseline; in the model, it is specified as a set of positive and negative impulses. The accent command, on the other hand, is relevant to the realization of accents and is specified as a step function. The phrase and accent commands are smoothed separately by their respective control mechanisms, which are approximated by a critically-damped second-order linear system. The outputs of these control mechanisms, which are in the model referred to as the phrase and accent components, are ultimately combined – specifically, the accent components are superimposed onto the phrase components – to generate surface F0 contours. Figure 2.1 which was adopted from Fujisaki (2003) shows how an F0 contour is synthesized using an example of Japanese declarative sentence.

The phrase command of this model is a partial instantiation of the *register*-control hypothesis, as it reflects the control of pitch register floor. As shown in the second panel of Figure 2.1, the accent components are superposed onto the phrase component, which makes the phrase component analogous to the register floor. Yet, as in the soft template model, the phrase command is different from the register in the current study: the phrase command specifies where an F0 contour should be located and what it should look like, but it does not specify the overall F0 *space* in which the targets are located. Similarly,

the accent command reflects the control of pitch targets, but it is not necessarily identical to the targets of the current study, as the accent command is modelled as a step function, which not only has information on the target value but also the duration of the target.

Therefore, both aspects of target and register control are incorporated in the command-response model, yet it exhibits some important differences from the target and register control put forward in this dissertation.

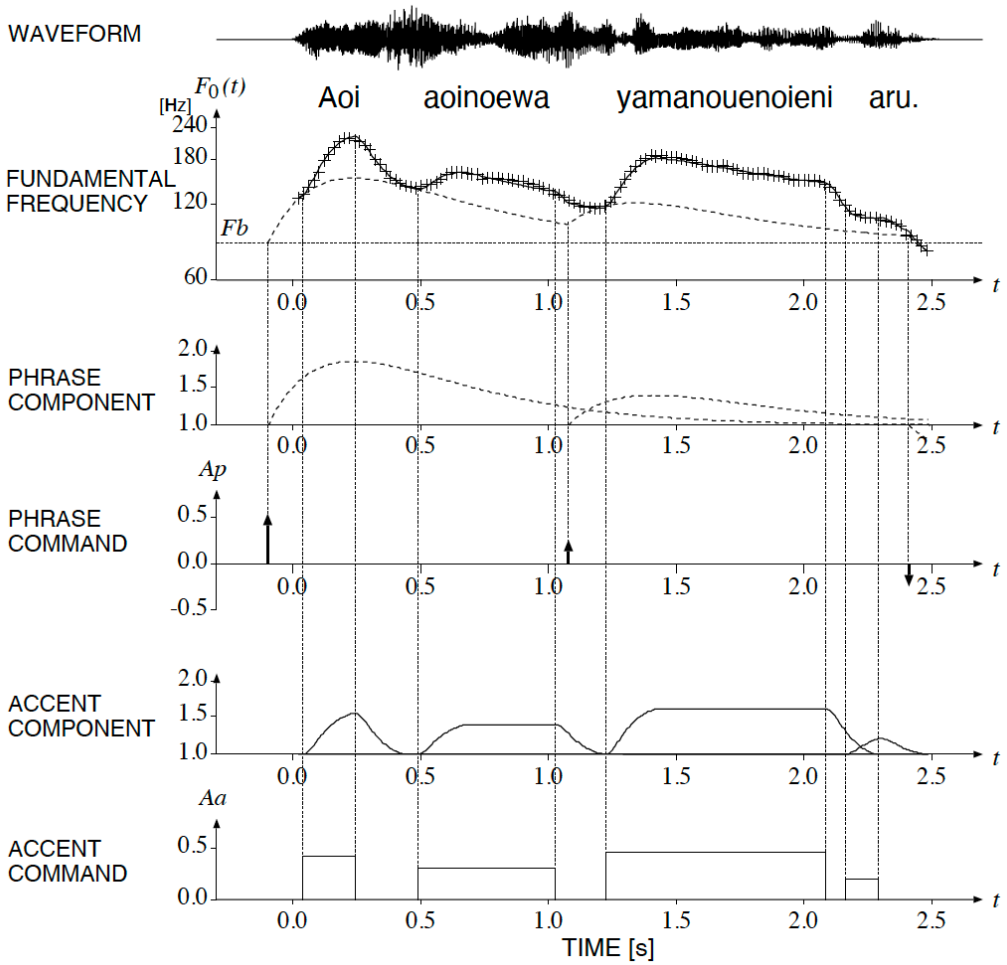

Figure 2.1. An example of F0 contour generation in the command-response model. The figure is adopted from Fujisaki (2003). The first panel presents a Japanese declarative sentence that is modeled in this example and its waveform. The phrase command (the fourth panel) is specified as positive/negative impulses (up/down arrows), and the accent command (the sixth panel) is specified as a step function. Each of these commands are smoothed by their respective control mechanisms, which result in the

phrase and accent components in the third and fifth panels, respectively. They are ultimately combined to generate a surface F0 contour as in the second panel.

2.1.5 PENTA model

The Parallel Encoding and Target Approximation (PENTA) model (Xu, 2005), which developed from the Target Approximation (TA) model (Xu & Wang, 2001), exemplifies both *target* and *register*-control hypotheses. The main idea of the PENTA model is that various communicative functions such as lexical, syntactic, or pragmatic information are encoded in F0 by specifying one or more control parameters of the model through its own encoding scheme. The control parameters of the model are pitch target, pitch range, articulatory strength, and duration, and they are referred to as melodic primitives. See Figure 2.2 for the schematic representation of the mechanism of the PENTA model, which is adopted from Xu (2005).

The TA/PENTA models consider both pitch target and range as the parameters that are actively controlled by speakers. Crucially, the pitch targets in these models refer to the *underlying* targets, not the *surface* targets that correspond to the turning points in the F0 contours (i.e. F0 peaks and valleys). Xu and Xu (2005) stated that the pitch targets in the PENTA model refer to "the articulatory goals that are *covert*" which may or may not correspond to "the *actual* peaks or valleys in surface F0 contours". This is to some extent similar to the general sense of pitch targets adopted in this dissertation, which refers to the abstract, cognitive representations of what speakers want their F0 to be that may or may not match the F0 peaks and valleys found in the surface contours.

These underlying targets are then continuously and asymptotically approximated from the onset to the offset of the syllable within a specified pitch range. The pitch range

parameter is similar to the pitch register of the current study in that both concepts specify the space that the F0 values can be realized at a given point in an utterance, contrary to the phrase curve of the soft-template model or the phrase command of the command-response model. Yet, it is not completely in line with the *register*-control in this dissertation, as the pitch range in the PENTA model simply delimits the space that the targets can be realized, but the control of pitch register in the current study is the main factor that can drive F0 variations.

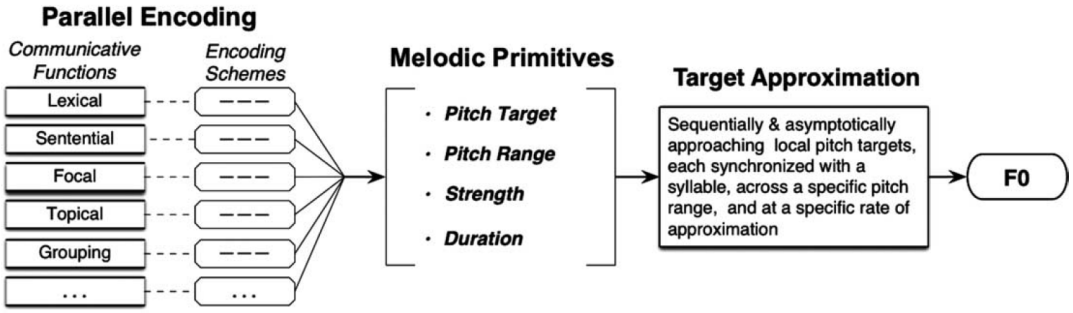

Figure 2.2. A schematic summary of the PENTA model, which is extracted from Xu (2005). Various communicative functions are encoded in F0 through their own encoding schemes, by controlling four melodic primitives in the middle of the figure. These primitives operate to sequentially and asymptotically approximate the underlying pitch targets which produce variations in F0.

The conceptual TA and PENTA models were also evaluated quantitatively in Prom-on et al. (2009), who introduced the quantitative Target Approximation (qTA) model. The model was tested on Mandarin Chinese and English data; the parameters of the model were derived by using the analysis-by-synthesis optimization algorithm. In both languages, the model-generated synthesized F0 contours exhibited low RMSE when they were compared to the empirical F0 contours, and the two contours were found to be highly correlated. In the perception experiment, listeners correctly identified tone

(Chinese), stress (English), and focus (Chinese, English) in the synthesized F0 data, and they judged the synthesized pitch contours to be natural.

2.1.6 Summary

Overall, the F0 models reviewed in the current section can be categorized in the following way given our main hypotheses of pitch control examined in this dissertation. Specifically, in our hypotheses, pitch targets are embodied as F0 *levels* and pitch register is understood to be the notion of *space*.

(i) *target*-control hypothesis: AM model (allow *register*-control)

(ii) *register*-control hypothesis: grid model

(iii) both: soft-template, command-response, PENTA models

In the AM model, different F0 peaks and valleys were considered to arise from the control of pitch targets, although it did allow some forms of register control. On the other hand, in the grid model, pitch register (grid lines) was the key factor that generates F0 variations. While the concept of pitch register/range has been discussed in many other F0 models, it played a more central role in the grid model as it was under active speaker control and results in different F0 peaks and valleys.

In the models in (iii), the control of both targets and register jointly contributed to the F0 variations, although they differed from the target and register control discussed in this dissertation. Specifically, unlike pitch register in the current study, which defines F0 space with a combination of at least two out of three of floor, ceiling, and span parameters, the phrase curve of the soft-template model and the phrase command of the

command-response model simply provided information about the overall F0 shape and direction. Moreover, the concepts that are analogous to pitch targets in these models were not necessarily the high and low pitch *levels* as assumed in this dissertation but were the representations that span for a certain amount of time with a specification on the target value as well as how that target is achieved.

In this dissertation, specifically in Chapter 4, the gestural model of F0 control is proposed and evaluated with the data collected from the experiment (Chapter 3). The model allows for a more direct comparison of the two main hypotheses of pitch control (*target* vs. *register*), which would further our understanding on the speakers' control system of F0.

One thing to note is that the previous F0 models were applied widely to both tone and intonation languages. For example, while the PENTA model was originally developed from the research on the tone language – i.e. Mandarin Chinese, it was extended to the intonation language; for instance, Xu and Xu (2005) and Prom-on et al. (2009) were able to successfully model F0 trajectories of English sentences with the PENTA model. In addition, the command-response model started from the pitch accent language – i.e. Japanese, but it was also applied to other languages including English, German, Greek, Korean, Polish, and Spanish (Fujisaki, 2003). Thus, it can be assumed that the core ideas of the F0 models are generalizable to different types of languages, rather than being restricted to a specific language that the model is developed from. In this sense, although the current F0 model is tested on the intonation language (i.e. English), it is assumed that in principle, the same control mechanism would apply to other types of languages as well.

2.2 Empirical F0 phenomena

This section introduces some empirical F0 phenomena and discusses how they have been (or can be) accounted for under our hypotheses of pitch control – i.e. *target* vs. *register* control. The first two phenomena are (i) *downstep* and (ii) *declination*, which are downward pitch movements across phrases or utterances that have been widely discussed in the literature. The other two phenomena are more directly relevant to the questions investigated in the production experiment of this dissertation (Chapter 3): they are the (iii) *pre-planned* pitch control with respect to sentence length and (iv) *adaptive* pitch control according to F0 perturbations in auditory feedback.

2.2.1 Downstep

Downstep is commonly described as the lowering of a high tone (H) after a low tone (L)[3]; for example, in the HLH tone sequence, the second H tone is realized at a lower pitch than the first H tone due to the intervening L tone. It is also not just a single H tone that is affected by the L tone, but all subsequent H tones are affected until reaching the end of a prosodic unit where the downstep is blocked – i.e. thus, the new ceiling is set for the H tones, resulting in a "terracing" effect.

Downstep was first studied extensively in tone languages (specifically, African

[3] Downstep illustrated here is an "automatic" downstep. There is also a "non-automatic" downstep, in which no conditioning factor (i.e. the intervening L tone) is present in the surface tonal string. For instance, the "non-automatic" downstep refers to the case where the second H tone is lowered compared to the first H tone in the HH sequence. It is usually considered that a floating or historically lost L tone triggers "non-automatic" downstep (see Connell, 2001).

tone languages) and then extended to intonation languages. The description of downstep phenomenon dates back to Christaller (1875) and Ward (1933), in their discussions of Fante (Kwa, Ghana) and Efik (Benue-Congo, Nigeria) grammars, respectively; yet, the current understanding of the nature of downstep was first offered in Winston (1960) in his work on Efik and was further improved with the development of the autosegmental phonology in the 80s and 90s (Connell, 2011). Downstep was first extended to non-tonal languages in Pierrehumbert (1980), where she accounted for the downward pitch movements found in English as a successive lowering of the pitch accents. In non-tonal languages, however, it has been debated whether the effect of downstep can be separated from declination (a global downward pitch movement), which will be further discussed in the next section.

Downstep is a "phonological" phenomenon, which is triggered by a low tone or tonal sequence (for example, an HL sequence in Japanese or a bitonal pitch accent in English) and is conditioned by a lexical, morphological, and syntactic structure of the utterance. Downstep effect is commonly explained with the *register*-control hypothesis; pitch register is shifted downwards with the occurrence of the downstep trigger, and it causes the lowering of subsequent tones until reaching the end of a prosodic unit or encountering another downstep trigger. Figure 2.3 shows the schematic illustrations of downstep that were introduced in the previous studies: as in both (a) and (b), downstep effects are modeled as the register shifting downwards. Note that it is the register ceiling and floor that is controlled, but not the span[4]. The *register*-control approach was adopted

[4] The register span, however, can vary when the downstep affects L tones differently from H tones (i.e. asymmetric effects of downstep on H and L tones). Connell and Ladd (1990) in fact pointed out that the

in many studies, such as Ladd (1983, 1990), Inkelas and Leben (1990), Clements (1990), and the later models of Pierrehumbert (Pierrehumbert & Beckman, 1988). Researchers have used a metrical tree structure (Ladd, 1990) or proposed a phonological feature such as [downstep] (Ladd, 1983) or an autosegmental tier called "register tier" (Inkelas & Leben, 1990) to formally describe the downstep effect.

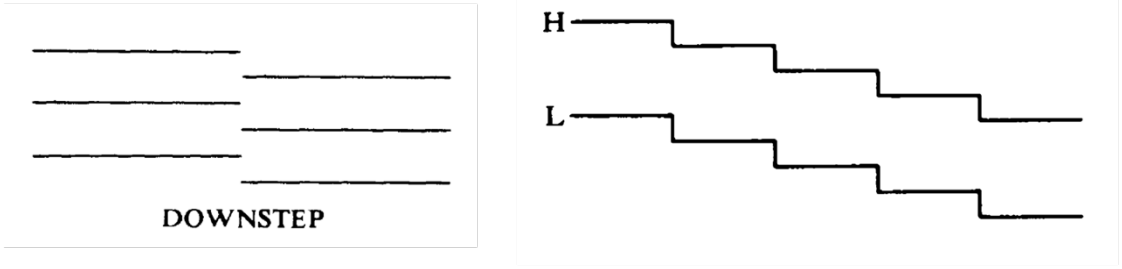

Figure 2.3. Schematic representations of downstep introduced in the previous studies. (a) is from Clements (1979); the top and the bottom lines represent register ceiling and floor, respectively, and the middle line shows the midpoint of the register. (b) is from Connell and Ladd (1990), which shows the realizations of H and L tones under the influence of downstep.

The *target*-control hypothesis can also account for the downstep phenomenon. In this hypothesis, the downstep rule is considered to apply on the individual tones; thus, the value of a given tone is computed with respect to the value of the preceding tone, in a way that the current tone is scaled lower relative to the previous tone. This idea was adopted in the Pierrehumbert's earlier model, such as Pierrehumbert (1980) and Liberman and Pierrehumbert (1984).

---

downstep is considered to be a *lowering* of the overall register in the Clements and Ladd models, while it is a *narrowing* of the register in the Pierrehumbert & Beckman model. In the latter case, the midpoint of the register remains constant and only the ceiling steps down.

2.2.2 Declination

Declination is defined as "a gradual modification (over the course of a phrase or utterance) of the phonetic backdrop against which the phonologically specified local F0 targets are scaled" (Connell & Ladd, 1990); "a titling of the graph paper" is the metaphor from Pierrehumbert (1980) that captures the declination effect. This phenomenon was first identified by Pike (1945), who reported the general tendency of F0 to "drift" down over the course of the sentence, and the term "declination" was coined by Cohen and 't Hart (1967). Unlike downstep, declination is considered to be a "phonetic" effect that refers to a continuous, long-term decline of F0 over the course of the phrase/utterance. See Figure 2.4, which was adopted from Ladd (1984) and illustrates declination effect in a hypothetical tone language; the range of F0 that is available for pitch targets (H, M, L tones) becomes gradually *lower* and *narrower,* such that the phonologically identical tones are realized in a different F0 depending on where they are in the utterance.

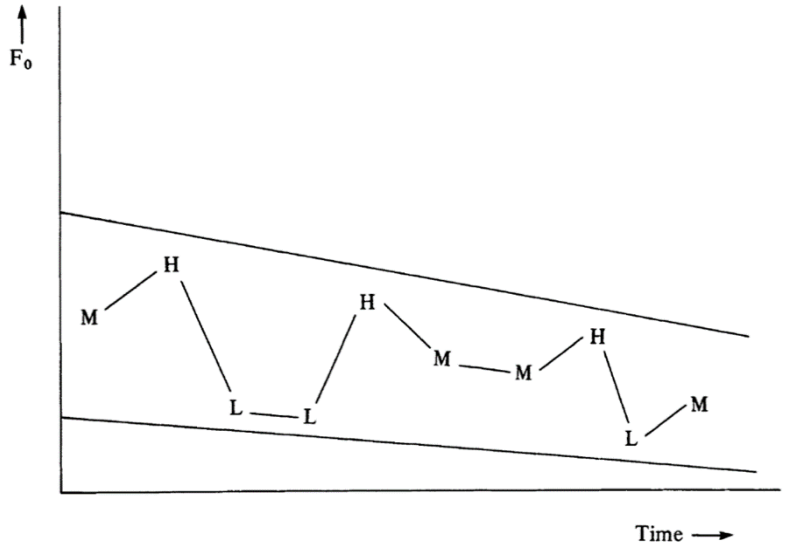

Figure 2.4. A schematic representation of declination extracted from Ladd (1984). The horizontal lines show the register ceiling and floor that exhibit the declination effect in a hypothetical tone language utterance.

Declination has been under debate on various aspects, the first of which is what causes declination. Specifically, previous studies have disagreed on whether declination is driven by a physiological mechanism or is under speaker control. For example, researchers such as Lieberman (1966) and Collier (1975) argued that the declination arises from the drop in the subglottal air pressure. Similarly, Maeda (1976) maintained that the declination is driven by a tracheal pull in addition to changes in subglottal air pressure. On the other hand, Ohala (1978) viewed declination not as a physiologically-driven automatic process, but more as a linguistic process which is actively controlled by speakers.

The second debate on declination is whether it needs to be considered as a distinct F0 phenomenon (i.e. a separate control parameter), or whether it is redundant with other downward F0 phenomena such as downstep. Liberman and Pierrehumbert (1984), in particular, argued that the declination effect can be explained by downstep and final lowering (an additional lowering of pitch at the end of the sentence), since F0 peaks found in the contours could be explained only with the combination of two effects. Recall that in Section 2.1.1, the Liberman and Pierrehumbert model only had a downstep parameter, but no declination parameter. This perspective, however, was later revised by Pierrehumbert and Beckman (1988), who argued that both declination and downstep are needed to model Japanese intonation.

Experiments were also carried out to find concrete evidence of declination. For example, Prieto et al. (1996) examined the scaling of F0 peaks in Mexcian Spanish downstepping contours. They hypothesized that if declination is present, a greater F0 reduction would be observed when the temporal distance (i.e. the number of syllables)

between the two accents is increased; if, however, only downstep matters, the amount of F0 reduction would be same regardless of the changes in the temporal distance. The results showed that the time-dependent lowering was almost absent in their data (i.e. no effect of temporal distance), supporting Liberman and Pierrehumbert (1984)'s argument that F0 downtrend can be exclusively explained by downstep.

As pointed out in the previous studies, declination effect is difficult to distinguish from downstep in intonation languages (Ladd, 2008), yet the two phenomena could be more readily separated in tone languages. Specifically, if any type of lowering is found in sentences that are composed of the same tones (e.g. all Hs or Ms), it can be attributed to declination, as there is no downstep trigger. This idea was tested in Shih (2000), who examined sequences of H tones with varying numbers of syllables in Mandarin Chinese. The results showed a decline of F0 over the utterance, providing evidence that the declination is an F0 phenomenon that is distinct from downstep.

If we consider declination as a separate phenomenon, the question that follows is how to model its effect. Under the *register*-control hypothesis, the declination effect can be understood to arise from the control of register floor and span. Specifically, it can be modelled as a gradual and constant lowering of the floor as well as the compression of the span, especially given that the range becomes smaller and narrower towards the end of the phrase/utterance as shown in Figure 2.4. (cf. The same effect can be derived if we assume the control of ceiling and span.) Instead of a linear decay of the floor, alternative forms of register control are possible. For instance, Shih (2000) proposed exponentially decaying declination model based on the empirical data that showed a faster declination rate at the beginning of the sentence and then slowed down later in the sentence; on the

other hand, Fujisaki (1983) modeled declination effect with a phrase curve, which had a form of an initial rise and a gradual, asymptotic decline (Section 2.1.4). It would not be impossible to model declination under the *target*-control hypothesis as well, although it must be assumed that the global F0 fall is considered when speakers compute the target value for each individual tone.

2.2.3 Sentence-initial pre-planned F0 control

Related to F0 declination, studies have examined whether speakers vary the initial F0 peak of a sentence according to the length of the sentence. Specifically, the studies tested whether speakers raise the initial F0 when they produce a long utterance. Since F0 declines over the course of the sentence (declination effect), speakers may start from a higher F0 in longer sentences, in order to avoid reaching the bottom of one's register before the utterance ends.

The correlation between sentence length and utterance-initial F0 peak has been therefore investigated in a variety of languages, including both intonation and tonal languages, but the studies have found mixed results. For instance, a significant correlation between initial F0 and sentence length was found in English (Cooper & Sorensen, 1981), Dutch (van Heuven, 2004), Danish (Thorsen, 1980), Swedish (Bruce, 1982), Mandarin Chinese (Shih, 2000), and Yoruba (Laniran & Clements, 2003), while no significant effect was observed in English (Liberman & Pierrehumbert, 1984), Dutch (van den Berg et al., 1992), Italian, Portuguese, and Spanish (Prieto et al., 2006), Mexican Spanish (Prieto et al., 1996), and Mambila (Connell, 2003, 2004). Even within the same language, the correlation results differed; for example, in English, Cooper and

Sorensen (1981) found a significant effect of utterance length on the initial peak height, while Liberman and Pierrehumbert (1984) found little effect. Similarly, in Dutch, van den Berg et al. (1992) showed that F0 values of the initial accents do not increase with the total number of accents in the utterance, while van Heuven (2004) found that the size of the first downstep is proportional to the number of items in the list that speakers have to produce.

There are also studies which found a significant correlation between the initial F0 peak and the length of the first constituent. One such example is Ladd and Johnson (1987); they recorded two native speakers of English and found that the height of the first accent peak is affected by the length of the sentence-initial constituent, but not so much by the length of the entire sentence. The similar result was found in German (Fuchs et al., 2013) and Wenzhou Chinese (Scholz & Chen, 2014), providing evidence that speakers may not consider the length of the whole sentence in their pitch control.

These inconsistent results altogether led researchers to believe that the initial F0 raising with respect to sentence length is a speaker-optional mechanism. In particular, borrowing the terminology of Liberman and Pierrehumbert (1984), it is generally agreed that the initial F0 control is part of "soft" preplanning, which is the preparation that speakers may freely choose to make, in contrast to "hard" preplanning, which is the essential preparation that speakers should accomplish before utterance initiation.

Regarding the control parameter, most of the studies mentioned above seemed to have assumed the *target*-control hypothesis, as they mainly examined utterance-initial F0 peak for the sentence length effect. In other words, the underlying assumption is that speakers would vary the utterance-initial H target according to sentence length. This,

however, led to inconsistent results. The current study thus approaches this question from a different perspective that speakers can indeed vary pitch *register* according to sentence length. In the experiment, participants were instructed to produce sentences that vary in length – particularly, the length of the subject phrase. The F0 values of the sentence-initial peak, valley, and the range between the two variables were examined to test whether speakers vary the whole tonal space (instead of just a single H target) according to sentence length. If speakers control pitch register as in the *register*-control hypothesis, we may observe a higher initial F0 peak, lower F0 valley, and/or wider F0 range.

2.2.4 Sentence-medial adaptive F0 control

The other question examined in the production experiment is whether speakers respond to changes in utterance length that are made after utterance initiation. As far as I am aware of, no studies have examined sentence-medial adaptive control of F0 in response to the changes in the length and content of the utterance. Therefore, in this section, I instead introduce the related studies which investigated the speakers' ability to adapt to the pitch-shifted auditory feedback using a perturbation experimental paradigm.

In these studies, participants were instructed to produce a single vowel, syllable, or sentence, and while producing them, they heard back their production with F0 perturbations – i.e. the pitch of their voice was either raised or lowered compared to the original production. Therefore, there was a mismatch between their actual production and the perception of their production in terms of F0. Most speakers exhibited

compensatory responses such that they produced F0 in the direction opposite to the shifted F0, although some showed following responses by producing F0 in the same direction to the perturbations. These results provided evidence that speakers are sensitive to information presented during production and further adjust F0 according to that information. Below, I summarize the findings of several studies on F0 perturbation and adaptive control.

First, studies have examined the effects of F0 perturbation during the production of a single vowel. For instance, Burnett et al. (1998) had participants produce a vowel /a/ for 5s, in which the onset of the vocalization activated auditory feedback that increased in pitch. In particular, participants were instructed to ignore any changes in the feedback, but to maintain their production as equally as possible throughout vowel phonation. Both compensatory and following responses were observed, although the former significantly outnumbered the latter. Jones and Munhall (2000) also investigated how speakers respond to pitch-altered auditory feedback during a vowel production task. Unlike Burnett et al. (1998) who altered F0 abruptly after the utterance was initiated, Jones and Munhall (2000) shifted F0 gradually to make changes less perceptible to participants. Participants exclusively exhibited compensatory responses, and interestingly, they continued to respond to F0 perturbations even after the perturbations were removed (i.e. when they heard the normal feedback in which the pitch was not altered).

Second, studies found evidence that speakers respond to pitch-shifted auditory feedback when they are producing a syllable. Natke and their colleagues (Donath et al., 2002; Natke et al., 2003; Natke & Kalveram, 2001) conducted pitch-altered perturbation

experiments with German speakers, where they were instructed to produce a nonsense word [ˈta:tatas] while hearing pitch-altered feedback. As in the vowel production, participants also showed compensatory responses while producing a syllable. Similarly, Jones and Munhall (2002) and Xu et al. (2004) found that speakers respond to pitch-altered auditory feedback even in tonal languages, where F0 was used contrastively.

Lastly, it was found that speakers are responsive to pitch-shifted auditory feedback when producing sentences, where F0 conveys sentence type or pragmatic information. In Chen et al. (2007), English speakers participated in two experiments where they had to produce a question and a sustained vowel, hearing the pitch-transformed feedback. They showed both following and compensatory responses in both tasks, although the response magnitude was larger in the question production task. Patel et al. (2011) examined the effects of pitch-altered auditory feedback on sentences with narrow focus, and here also, participants adjusted F0 in response to the shifted feedback.

This body of literature provides ample evidence that speakers monitor information that is presented during production and further have ability to adapt to that information almost real-time. This is promising for the current study, as speakers may also adjust F0 parameters according to the changes in sentence length that are made after utterance initiation. Related to our hypotheses of pitch control, previous studies on pitch-shifted auditory feedback seemed to have mainly assumed that speakers control pitch *targets*. The studies examined how F0 values of a vowel, syllable, or sentence vary with respect to pitch-shifted auditory feedback rather than how the overall tonal space is varied; in fact, in the vowel or syllable production tasks, the *register*-control hypothesis could not even be tested.

In the current experiment, there is one condition where the length of the sentence is fully provided before utterance initiation (thus, no changes in length sentence-medially) and another condition where length changes after the start of production. If speakers vary pitch register, speakers would adjust F0 parameters – i.e. F0 peaks, valleys, and ranges – differently in these two conditions. Specifically, it is expected that speakers would raise the ceiling and/or broaden span if they encounter a newly presented part of the sentence.

2.2.5 Summary

This section introduced some empirical F0 phenomena and discussed them in relation to the *target* and *register*-control hypotheses. To summarize, downstep effect can be considered to arise from register shifts under the *register*-control hypothesis, whereas in the *target*-control hypothesis, it is assumed that speakers lower a current pitch target considering the preceding targets. The global decline of F0 over the course of the phrase/utterance (i.e. declination) is more readily associated with the control of register – specifically, the register floor and/or span. In Chapter 4, the two alternative accounts of downstep (i.e. *target* vs. *register*) are directly compared via computational modeling; the declination effect is modeled through the register floor parameter in the current F0 model.

Regarding the speakers' pre-planned and adaptive F0 control, previous studies have mostly assumed that speakers control pitch *targets* to produce variations according to the initial sentence length and in response to the F0 perturbations. For instance, in the previous studies, the utterance-initial F0 peak was predominantly examined to find out

the effect of sentence length in the speakers' initial F0 control. In the current experiment, which is introduced in Chapter 3, the possibility of speaker's *register*-control is also explored, by examining not only F0 peaks but also valleys and ranges.

2.3 Summary of background

In this chapter, I have reviewed the conceptual/computational F0 models proposed in the literature and some empirical F0 phenomena, especially from the lens of our main hypotheses of pitch control – i.e. *target* and *register*-control hypotheses. In Section 2.1, we have identified the model that is mainly built upon the *target*-control hypothesis (AM model) and the *register*-control hypothesis (grid model). There were also models (soft-template, command-response, PENTA models) which incorporated both aspects of *target* and *register* control, yet their conceptualizations of the *target/register*-control somewhat differed from what is assumed in this dissertation.

No studies, however, have tried to directly compare the two alternative hypotheses in terms of how well they account for F0 variations observed in the empirical data. For instance, the empirical F0 phenomenon of downstep could be modeled either with the *target* or *register*-control hypothesis, as we have seen in Section 2.2.1. All five models introduced in Section 2.1 thus would be able to model downstep, but we do not know which F0 model – more specifically, which F0 control hypothesis – better explains the downstep phenomenon. This comparison, however, is important, as the model that better accounts for the data can be understood as a better manifestation of the speakers' cognitive F0 control mechanism.

In this sense, this dissertation aims to examine and more directly compare the two hypotheses of F0 control through production experiment and computational modeling. In the experiment (Chapter 3), I investigate how speakers vary F0 parameters according to the initially planned sentence length and changes in the length that are made after utterance initiation. This would complement previous studies, by specifically exploring the possibility of *register*-control, which has been neglected in the literature (Sections 2.2.3, 2.2.4). The data collected from the experiment are used to test the dynamical model introduced in Chapter 4. The models that exemplify *target*-control and *register*-control hypotheses are compared in terms of how well they fit the empirical data. The results from the experiment and modeling would ultimately provide evidence on what it is that speakers control (i.e. *targets* vs. *register*) to produce variations in F0.

CHAPTER 3
PRODUCTION EXPERIMENT

3.1 Introduction

In Chapter 1 and Chapter 2, I proposed two possible hypotheses of F0 control – *target* vs. *register*-control – and illustrated how they are reflected in the F0 models proposed in the literature and how they can explain some empirical F0 phenomena such as downstep and declination. In Chapter 3 and Chapter 4, the two hypotheses are more directly compared in terms of their validity through a production experiment (Chapter 3) and a modeling study (Chapter 4). The main goal of these studies is to find out which hypothesis (*target* vs. *control*) better accounts for the empirical data and thus better reflects the speakers' F0 control mechanism.

In this chapter, I introduce a sentence production experiment that is designed to examine two aspects of speakers' F0 control: the (i) pre-planned and (ii) adaptive F0 control. In particular, the experiment examines how speakers control F0 parameters (i) according to the initial sentence length – i.e. the F0 control set before utterance initiation – and (ii) in response to the unanticipated changes in sentence length – i.e. the adaptive F0 control during production. There are two main objectives for this chapter: the first is to find out how speakers vary F0 parameters according to the experiment manipulations (i.e. the initial sentence length and changes in the length), and the second is to identify what it is that speakers control – *target* vs. *register* – to produce these F0 variations (i.e. what the observed variations inform us about their F0 control).

To test the speakers' pre-planned/initial F0 control, I manipulated the number of

noun phrases (NPs) in the subject phrase, such that the sentence had one, two, or three subject NPs. To test the speakers' adaptive F0 control, a novel experimental paradigm was developed in which the parts of the sentence were delayed until after speakers initiated production – i.e. the non-initial NPs of sentences that had two or three subject NPs were presented not at the beginning of the trial but immediately after participants started production. In this case, participants had to incorporate the delayed phrases into their ongoing utterance as smoothly as possible.

Two sets of analyses were conducted. The first set of analyses examined how the values of F0 peaks/valleys/ranges of each NP as well as changes in the values across NPs differ by the initial sentence length and the occurrence of delayed NPs. This was to find out how speakers vary F0 parameters according to the experiment manipulations.

The second set of analyses examined the variance and correlation of F0 measures and compared the condition-prediction models, with the goal of identifying the control parameter (*target* vs. *register*). Crucially, these analyses required an assumption where the measures of F0 peaks represent H pitch targets, F0 valleys to represent L targets, and F0 ranges to reflect the register span. As discussed in Section 1.3, F0 values of peaks and valleys are the reasonable estimates of pitch targets, since a target undershoot is less likely in speech with a relatively moderate tempo. In addition, F0 range is the decent estimate of register span; although F0 peaks and valleys may not be at the edges of the register (i.e. thus, the F0 range measure is underestimating the actual register span), the range would still be highly correlated with the span. Note that the F0 peaks and valleys could also be considered to reflect the register ceiling and floor, but here, I consider them only as the representations of pitch targets and the range to reflect the register.

For a brief summary of the results, most F0 variables showed a significant effect of initial sentence length and delayed stimuli presentation, providing evidence for the pre-planned and adaptive F0 control. Regarding the examinations on the F0 control mechanism (i.e. *target* vs. *register*), the results found evidence for the *register*-control. Below, I present specific hypotheses and predictions of the analyses conducted in this chapter. The rest of the section is organized as follows. Section 3.2 describes the experiment design and data processing and analysis methods. Section 3.3 presents the results, and Section 3.4 discusses them.

3.1.1 Hypotheses and predictions

The first hypothesis that is tested in the experiment is that speakers vary F0 parameters according to the initial sentence length – i.e. they make a pre-utterance F0 plan considering the initial utterance length. The motivation behind this hypothesis is that speakers would raise their initial F0 in longer sentences in order to avoid hitting the pitch register floor before reaching the end of the utterance (see Section 2.2.3). Figure 3.1 shows the comparisons that are made in the analyses and illustrates the predictions. As shown in (i), to examine the speakers' pre-planned F0 control, F0 measures (peaks, valleys, ranges) of the first NP were compared across conditions with different numbers of initial NPs. The prediction is that the F0 values of NP1 would significantly differ by the initial utterance length; for instance, F0 peaks would increase and/or ranges would expand in the utterances with more initial NPs.

The second hypothesis that is tested is that the speakers adjust their F0 control to adapt to the unanticipated changes in the length and content of the utterance. If this is

the case, the adjustment of F0 peaks/valleys/ranges from the first to the second NP in particular (as the delayed stimuli appear immediately after the utterance is initiated) would significantly differ between conditions in which the stimuli are delayed vs. that are not. See Figure 3.1-(ii), which shows the comparison of F0 peak adjustments from NP1 to NP2 between conditions with vs. without delayed stimuli. The logic behind this prediction is that if participants indeed adjust their control of F0 once they encounter delayed phrases, it should result in some F0 differences between conditions with vs. without delayed stimuli.

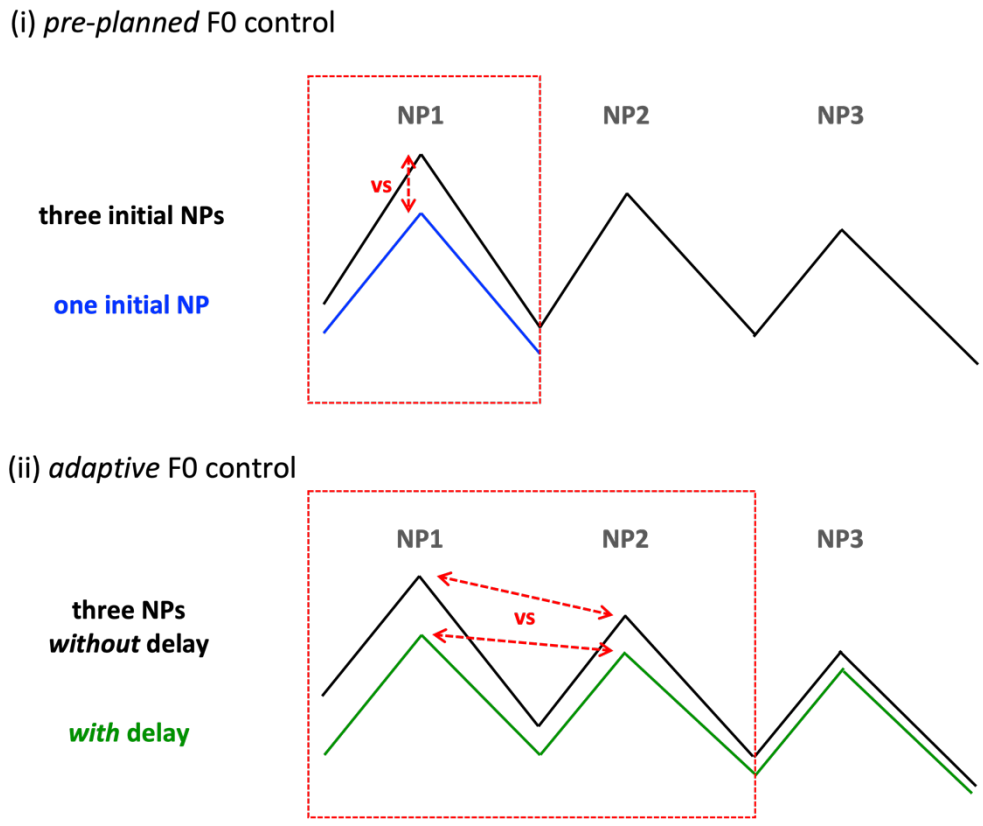

Figure 3.1. Schematic illustrations of the comparisons made in the analyses and the predictions. The solid lines represent schematic F0 contours. The red boxes mark the target regions of the investigation, and the red arrows indicate what is compared. Note that the figures show only the comparisons of F0 peaks, but the analyses were conducted on other F0 measures such as valleys and ranges.

Regarding the analyses of the speakers' control mechanism, specific predictions of the *target* and *register*-control hypotheses are presented in Table 3.1. In the (i) variance analysis, if the sum of the variances of F0 peaks and valleys is substantially larger than the variance of ranges, it would provide evidence for the *register*-control hypothesis, as this would show that the two variables are not independently controlled. If, however, the variances that are compared are almost equal, it would provide evidence for the *target*-control hypothesis. In the (ii) correlation analysis, a high correlation between peaks and valleys would provide evidence for the *register*-control hypothesis. (cf. Although the *target*-control hypothesis would essentially make the same prediction, the correlation is interpreted here as the evidence for *register*-control; I thus marked N/A for the *target*-control hypothesis in Table 3.1). In (iii) the comparison of condition-prediction models, if the AIC value of the model that has either F0 peaks or valleys as the predictor is lower than the model that has F0 ranges as the predictor, it would provide evidence for the *target*-control hypothesis. The contrary result – i.e. a lower AIC in the model with the ranges as the predictor – would lend support to the *register*-control hypothesis.

Table 3.1. Predictions of the F0 control hypotheses. The second and third columns show how each control hypothesis would predict the results of the analyses. As mentioned above, this study considers F0 peaks and valleys to be the indirect estimates of H and L pitch targets, and F0 ranges to be the estimates of register span.

		***target*-control hypothesis**	***register*-control hypothesis**
(i)	variance of F0 measures	σ^2(range) ≈ σ^2(peak) + σ^2(valley)	σ^2(range) ≪ σ^2(peak) + σ^2(valley)
(ii)	correlation of F0 measures	N/A	peaks and valleys within NP are highly correlated
(iii)	model comparison (using AIC)	range model > peak, valley models	range model < peak, valley models

3.2 Methods

3.2.1 Participants and experiment design

Thirteen native speakers of English (7M, 6F) with no speech or hearing disorders participated in the experiment. Participants read sentences in which the subject phrase was composed of one, two, or three conjoined noun phrases (NPs) and ended with the verb phrase (VP) *"live in the zoo"*. The lexical and phonological content of the conjoined NPs was carefully controlled such that each of them was composed of monosyllabic numeral ("*eight, nine*") + monosyllabic color (*"red, green, blue"*) + disyllabic animal with initial stress (*"llamas, rhinos, weasels"*). Table 3.2 presents sample stimuli in each experimental condition.

In sentences with multiple NPs, participants were instructed to connect them with the conjunction *"and"*. In the preliminary tests of the experiment, which were conducted on two native speakers of English and myself, various ways of connecting NPs were considered – for example, NPs were connected without the conjunction (e.g.

"*Nine green rhinos, eight red weasels, eight blue llamas live in the zoo"*) or inserting *"and"* only before the last phrase (e.g. "*Nine green rhinos, eight red weasels, and eight blue llamas live in the zoo"*). All participants of the preliminary tests confirmed that inserting *"and"* between every NP was most natural and easiest to produce, so that form was adopted in the experiment. All NPs were cued with the visual stimuli as in Figure 3.2; the VP was not visually cued, as it was repeated in every trial.

Table 3.2. Experimental conditions and sample stimuli. The stimuli had one, two, or three NPs in the subject phrase and ended with the VP "live in the zoo"; the subject NPs were connected with "and". The phrases marked in yellow were presented at the beginning of the trial, while those in green were presented after detection of utterance initiation (i.e. delayed).

NP1		**NP2**		**NP3**	**VP**
3NS. *3NPs + no-delayed stimuli*					
Nine green rhinos	and	Eight red weasels	and	Eight blue llamas	live in the zoo
3DS. *3NPs + delayed stimuli*					
Nine green rhinos	and	Eight red weasels	and	Eight blue llamas	live in the zoo
2NS. *2NPs + no-delayed stimuli*					
Nine green rhinos	and	Eight red weasels			live in the zoo
2DS. *2NPs + delayed stimuli*					
Nine green rhinos	and	Eight red weasels			live in the zoo
1NS. *1NP + no-delayed stimuli*					
Nine green rhinos					live in the zoo

For the sentences with multiple NPs, a condition was tested in which the visual stimuli that cued non-initial phrases were delayed until after detection of utterance initiation. In this condition, participants saw only one visual stimulus before production, and as soon as they started speaking, the remaining stimuli (phrases colored in green in Table 3.2) appeared on the screen. This condition is referred to as the *delayed stimuli* (DS) condition, as opposed to the *no-delayed stimuli* (NS) condition, where all NP stimuli were presented before production. Note that there was no three-NP condition in which the two NP stimuli were presented before production and one NP stimulus was

delayed; the reason for this relates to limitations in the control of the timing of delayed stimuli, which are discussed in Section 3.4.2. The sentences in the DS condition, therefore, were always cued with one initial stimulus and one (2DS) or two (3DS) delayed stimuli. Overall, three utterance lengths (1/2/3 NP) × two delay conditions (NS/DS) were tested in the experiment – i.e. 1NS, 2NS, 2DS, 3NS, 3DS; the DS condition could not be tested for the single NP stimuli (i.e. no 1DS).

There were nine blocks of 30 trials in each experimental session. In each block, there were five trials of 2NS, 2DS, 3NS, and 3DS, respectively, and ten trials of 1NS; the number of 1NS trials was doubled to compensate for the missing 1DS counterpart. These conditions were varied randomly from trial to trial. The numerals, colors, and animals were randomly selected in a way that ensured each word to appear at least certain number of times within a block; in particular, a total of 60 NPs occurred in each block, and they must have each numeral ("*eight, nine"*) more than 27 times, and each color ("*red, green, blue"*) and animal ("*llamas, rhinos, weasels"*) more than 17 times. In the trial with multiple NPs (i.e. 2NS, 2DS, 3NS, 3DS), the animals of the NPs were unique, although the numerals and colors could be same across different NPs.

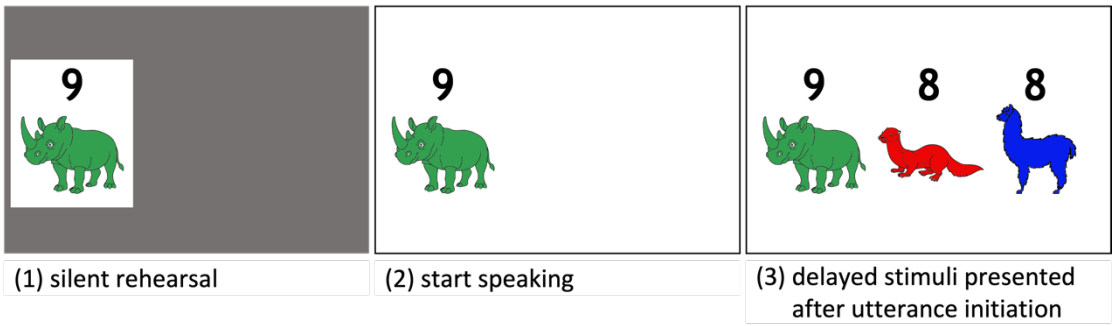

Figure 3.2. Presentation of a single trial. (1): The initial visual stimuli were presented with a grey background, and participants were instructed to silently rehearse the sentence. (2): After some periods of time, the grey background automatically changed to white, and participants could start speaking. (3): In the delayed-stimuli (DS) conditions, the visual stimuli that cued non-initial phrases appeared as soon as the utterance initiation was detected; participants should incorporate the delayed phrases into their sentences.

Participants were cued to the sentence as illustrated in Figure 3.2. In each trial, the initial visual stimuli appeared on a grey background as in Figure 3.2-(1). Participants were told that this is the preparation stage, and they were instructed to "silently rehearse" the sentence during this period – i.e. make a sentence in their head, without making sounds or moving their articulators. The time for the silent rehearsal varied by the number of initial stimuli; it was 2.7s for one initial stimulus, 4.4s for two initial stimuli, and 6.1s for three stimuli. These durations were derived from the average durations of the target sentences in the pilot data and additionally tested on two native speakers of English who were naïve about the experiment.

After these periods, the background automatically changed to white as in Figure 3.2-(2), which was the signal that cued participants to start speaking. In the DS conditions, the delayed stimuli (the visual stimuli that cued non-initial phrases) appeared immediately after utterance initiation was detected as in Figure 3.2-(3). The algorithm

for utterance initiation detection is introduced in the section below. Participants were instructed to incorporate the delayed stimuli smoothly into their ongoing utterance.

In order to prevent participants from putting contrastive focus on the parts of the NPs, which may affect the natural intonation of the sentence, participants were explicitly instructed not to emphasize any of the words in the utterance. Before the start of the experiment, participants did a practice session, which was composed of 24 trials that included all conditions (i.e. 1NS, 2NS, 2DS, 3NS, 3DS) and all numerals, colors, and animals. The purpose of the practice session was to make participants become familiar with the novel experimental design (delayed stimuli) and ensure that they read sentences naturally without any contrastive focus.

3.2.2 Detection of utterance initiation

The algorithm for detecting utterance initiation worked as follows. See Figure 3.3 below. As soon as the recording was initiated (Figure 3.3-(a)), a speech detection algorithm was applied at 1 ms intervals. The algorithm calculated the mean absolute amplitude of the preceding 100 ms of the acoustic signal. If the mean absolute amplitude of the signal in this 100 ms window was above a threshold value, that frame was considered to contain the onset of the utterance (Figure 3.3-(b),(c)), and the delayed stimuli appeared (Figure 3.3-(d)). Therefore, the fastest time that the delayed stimuli could appear (except for a short variable time for the image to be drawn to the screen, which is probably less than 20 ms) was the end of the frame that the signal went above the threshold. I used 0.015 as the amplitude threshold, which was determined through the preliminary tests of the experiment. During the practice sessions, the experimenter

monitored whether there were any trials that had delayed stimuli presented before the start of the utterance, to make sure that the threshold of 0.015 was appropriate for a given participant. Overall, the threshold and the detection algorithm worked well in all experimental sessions.

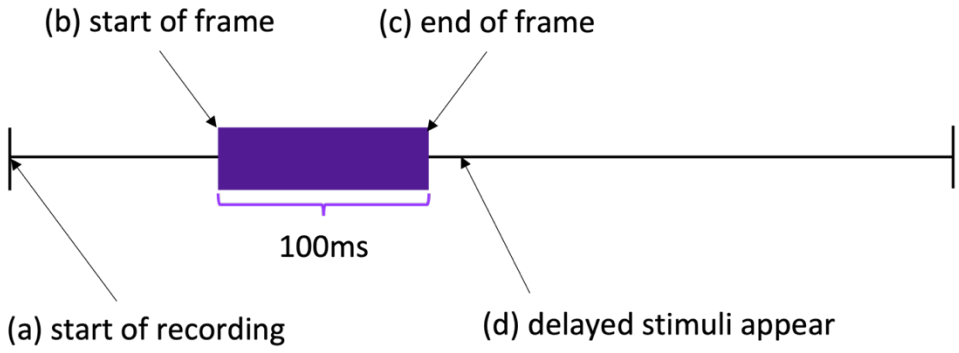

Figure 3.3. A schematic representation of utterance onset detection and delayed stimuli presentation. The horizontal line represents the flow of the acoustic signal. From the (a) start of the recording, 100ms length of the frame was constantly monitored at 1 ms intervals. The purple box indicates the frame that the mean absolute amplitude of the signal went above a threshold value (i.e. detection of utterance initiation), and (b) and (c) mark the start/end timepoints of the frame. The delayed stimuli were presented as soon as the frame was detected (d).

A post-hoc analysis on the timing of the delayed stimuli found that the stimuli appeared on average 86.4 ms after utterance initiation (as determined by the forced alignment). This means that the difference between the utterance onset identified by the forced alignment (which is explained in Section 3.2.3) and the end of the frame (i.e. the fastest time that the delayed stimuli could appear) was on average 86.4ms. Figure 3.4 presents the overall distribution of the differences. For DS trials, the utterance onset (occurs within the purple box in Figure 3.3) and the endpoint of the frame (Figure 3.3-(c)) were compared to identify the stimuli presentation error; when the end of the frame preceded the onset, it was marked as an error; see Section 3.2.3 for further details.

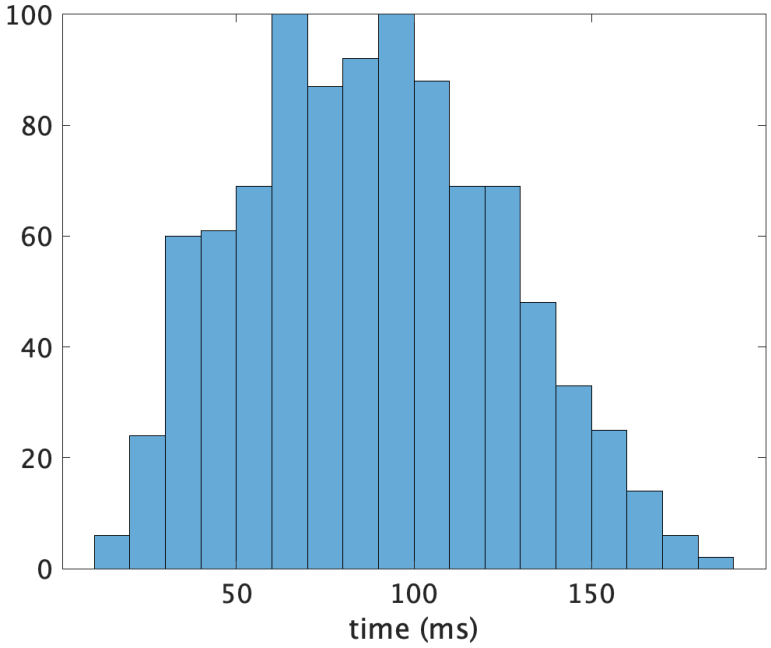

Figure 3.4. Distributions of differences between the endpoint of the frame that contained utterance initiation (Figure 3.3-(c)) and the utterance onset determined by the forced alignment. The x-axis is their differences in ms.

3.2.3 Data collection and exclusion

The experiments were conducted in a sound-attenuating booth. Participants wore a condenser microphone (AKG C520 headset), and acoustic data were collected at a sampling rate of 22050 Hz. Acoustic segmentation was carried out using Kaldi (Povey et al., 2011). For each participant, 10 randomly selected trials, which included all numerals, colors, and animals, were manually segmented and used to train monophone HMMs. A forced alignment was then conducted on all trials, and the alignments were manually inspected and corrected if necessary.

As mentioned in the previous section, for the trials in the DS conditions, the timepoint of the end of the frame that the utterance was assumed to have initiated was

compared with the onset of the utterance determined by acoustic segmentation. In four out of 3510 trials (0.1%), the start of the utterance followed the endpoint of the frame, which suggests that the delayed stimuli may have appeared before participants initiated the utterance. In addition, 11 out of 3510 trials (0.3%) had problems in data collection. These 15 trials were excluded from the analyses (see Table 3.4 for the distribution of these trials by participant), which left a total of 3495 trials (99.6%).

Due to the novelty of the experiment design, in which some visual stimuli were presented after the start of the utterance, participants may have produced disfluencies such as hesitations or speech errors. To identify these trials algorithmically, a mixed-effects linear regression was conducted on the durations of the words and between-word silence intervals (when present). At each word and silence location (e.g. in NP1, the numeral, the (possible) silence between the numeral and color, the color, the (possible) silence between the color and animal, the animal), a mixed-effects linear model was fit to the data with the experimental conditions (i.e. 1NS, 2NS, 2DS, 3NS, 3DS) as a fixed effect and the participant as a random intercept. At each location, datapoints whose standardized absolute residuals were larger than 3.09 (the 0.1/99.9 percentiles of a normal distribution) were considered as duration outliers. A total of 413 out of 3495 trials (11.8%) were excluded from the analyses, as they suggested the presence of disfluencies.

The cross-tabulations of trials by the occurrence of duration outliers found that the trials in certain conditions were more likely to exhibit signs of disfluencies. The number of trials with duration outliers was compared with the number of trials that did not contain any errors (i.e. no problems in data collection or stimuli presentation). As

presented in Table 3.3, the trials that had longer sentence stimuli (i.e. more subject NPs) were more likely to have duration outliers ($X^2(2, N = 3495) = 149.53, p < 0.001$). Within the trials with multiple subject NPs, the trials in which the stimuli were delayed tended to have more duration outliers than those without the delayed stimuli ($X^2(1, N = 2328) = 7.67, p < 0.01$).

Table 3.3. Cross-tabulations of trials by the occurrence of duration outliers. The top table shows the data grouped by sentence length (1NP vs. 2NPs vs. 3NPs). The bottom table shows the trials with two or three NPs, grouped by the delayed stimuli presentation (no-delayed vs. delayed stimuli). Each cell presents the number of trials with or without duration outliers for a given condition along with the percentage.

	1NP	**2NPs**	**3NPs**
trials with duration outliers	40 (3.4%)	143 (12.3%)	230 (19.8%)
trials without any errors	1127 (96.6%)	1021 (87.7%)	934 (80.2%)

	no-delayed stimuli	**delayed stimuli**
trials with duration outliers	162 (13.9%)	211 (18.1%)
trials without any errors	1002 (86.1%)	953 (81.9%)

After excluding the trials with duration outliers, there were two participants for whom the exclusions constituted more than 20% of that participant's data. Table 3.4 provides a summary of the problematic trials (problems in data collection and delayed stimuli presentation) and duration outliers for each participant. In particular, 75.2% of the data remained for one participant (PA12), and 69.7% of the data remained for the other participant (PA13). This low percentage of the remaining data suggests that these participants did not conform to the task instructions, either perhaps because they found

the task too difficult, or maybe they simply did not understand the instructions. The data from these two participants were thus excluded, and only the data of 11 participants were analyzed.

Table 3.4. The numbers of problematic trials and duration outliers by participant. The numbers in the data collection and delayed stimuli rows show the number of trials that had problems in data collection and the trials in which the delayed stimuli appeared before utterance initiation. Total 1 is the number of total trials (270) minus these two erroneous trials. Total 2 shows the number of remaining trials after excluding duration outliers. The percentage in total 2 is calculated by dividing the number of remaining trials by total 1 (trials without any errors).

	PA01	**PA02**	**PA03**	**PA04**	**PA05**	**PA06**	**PA07**
data collection	-	-	-	2	3	-	1
delayed stimuli	-	-	-	1	1	-	-
total 1	270	270	270	267	266	270	269
duration outlier	35	9	16	29	11	30	34
total 2	235 (87%)	261 (96.7%)	254 (94.1%)	238 (89.1%)	255 (95.9%)	240 (88.9%)	235 (87.4%)

	PA08	**PA09**	**PA10**	**PA11**	**PA12**	**PA13**
data collection	1	-	2	-	-	2
delayed stimuli	-	-	-	1	-	1
total 1	269	270	268	269	270	267
duration outlier	23	6	24	48	67	81
total 2	246 (91.4%)	264 (97.8%)	244 (91%)	221 (82.2%)	203 (75.2%)	186 (69.7%)

To summarize the data from the 11 participants, out of 2970 trials (270 trials x 11 participants), a total of 277 trials (9.3%) were identified to have problems in delayed stimuli presentation/data collection or contain duration outliers. Specifically, nine trials (0.3%) had problems in data collection; three trials (0.1%) were considered to have delayed stimuli presented before utterance initiation; and 265 trials (8.9%) contained duration outliers. Although it may seem that quite many trials were excluded due to the duration outliers, it is important to consider that the nature of the task – both the use of

multi-feature images and delayed stimuli – may have induced disfluencies more often than conventional utterance elicitation with read sentences. Overall, a total of 2693 trials (90.3%) were subject to subsequent analyses.

3.2.4 F0 processing

F0 trajectories were extracted using Praat as follows. First, participant-specific F0 range was identified by collecting a first-pass of F0 values from all trials of a given participant using a relatively broad gender-specific F0 range; for male speakers, the provisional pitch floor and ceiling were set as 40 and 200 Hz, while for female speakers, they were set as 100 and 320 Hz. The pitch range for a given participant was determined as 2.5-97.5% range of their F0 distribution obtained from all of the F0 values in the first-pass.

Second, besides pitch floor and ceiling, other Praat settings that are related to the post-processing pitch extraction algorithm (i.e. which determines the cheapest path through the pitch candidates) were tested in order to obtain more accurate F0 values. The parameters that were tested were *octave-jump cost* (the degree of disfavoring of pitch changes) and *voiced/unvoiced cost* (the degree of disfavoring of voiced/unvoiced transitions), which are specifically relevant to the F0 changes over adjacent frames. Different combinations of these parameter values were tested on the first 10 trials of each participant, the results of which were qualitatively assessed. The F0 values that were judged most accurate, considering the segmental properties (e.g. no F0 in voiceless sounds) were obtained from the combination of *octave-jump cost* as 0.7 (cf. standard value: 0.35) and *voiced/unvoiced cost* as 0.28 (cf. standard value: 0.14), which I adopted

for F0 extraction. For all other parameters, Praat default values were used. Using this set of Praat setting parameters and participant-specific F0 range, F0 data for all trials were extracted with a timestep of 5ms using the auto-correlation method.

F0 outliers were identified under two criteria: (a) the number of surrounding frames without an F0 value and (b) the F0 differences between the frames. For a given F0 contour, these two criteria were constantly applied until there were no more frames to remove. First, if a given F0 frame was surrounded by a sequence of frames without an F0 value, the value of that given frame was considered as an erroneous extraction of F0. Specifically, for a given frame with an F0 value, ten preceding frames and ten following frames were examined; out of these 20 frames, if more than 18 frames lacked an F0 value, the given frame was identified as an error. The red dots in panel (a) of Figure 3.5 indicate the frames identified as errors under this criterion.

Second, if the F0 of a given frame was preceded or followed by a large F0 jump, further inspections were conducted on the region surrounding that frame to identify an error. In particular, the threshold for a large jump was set as 22 Hz; to obtain this value, the absolute F0 differences between successive frames were collected from all trials of all participants, and 22 Hz was at 99.7% of the difference distribution. If the F0 difference between the two frames was more than 22 Hz × frame distance (e.g. frame1-frame2, the difference threshold is 22 Hz; frame1-missing value-frame2, the difference threshold is 44 Hz; frame1-missing value1-missing value2-frame2, the difference threshold is 66 Hz), that region was considered to potentially contain an F0 error.

In this case, in order to decide which frame (i.e. frame1 vs. frame2) to remove, I examined the general trend of F0 trajectory surrounding those frames. For example,

suppose that the F0 difference between frame1 and frame2 was larger than 22 Hz (e.g. the blue dots in Figure 3.5-(b)); I searched for the nearest frame that does not have an F0 value on the left side of the frame1 and the nearest frame without an F0 value on the right side of the frame2. These nearest frames defined the start and end points of the chunk of F0 trajectory that needed further investigation (e.g. the vertical red lines in Figure 3.5-(b)). I then calculated the mean F0 value of this region, excluding the F0 values of frame1 and frame2. That mean was compared to each of the F0 values of frame1 and frame2, and the frame that had the F0 value farther away from the mean was considered to be an erroneous extraction of F0 and thus discarded.

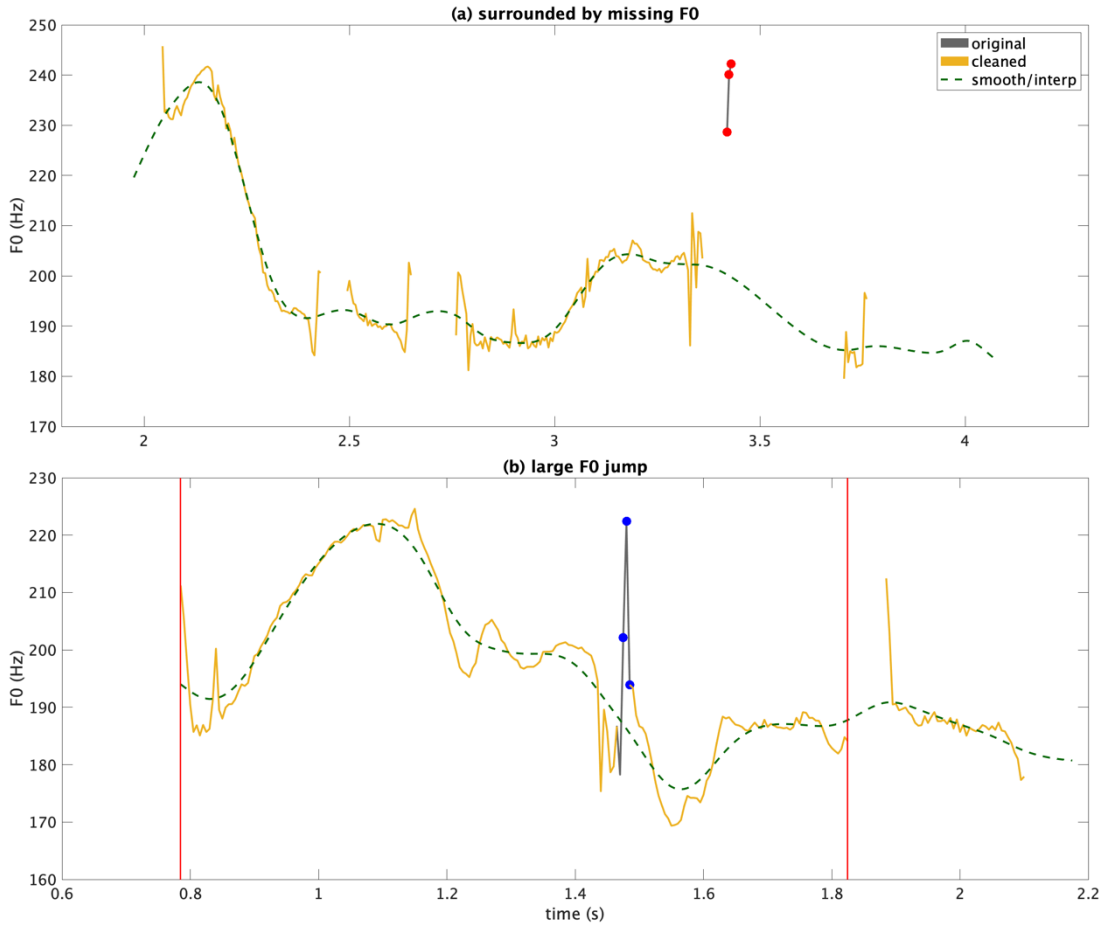

Figure 3.5. An example of F0 outlier detection. The (a) top panel shows the outliers surrounded by a sequence of frames without F0 values. The (b) bottom panel shows the outliers that exhibit a large F0 jump with adjacent frames. The problematic frames are marked as red and blue dots. The original raw F0 contour is plotted in grey, the cleaned contour after outlier removal process is plotted in yellow, and the smoothed and interpolated contour is plotted as a green dashed line. The red vertical lines in (b) mark the chunk of F0 trajectory that was used to determine which frame to be removed.

After removing F0 outliers, if for a given trial, the number of frames that had an F0 value was less than 50% of the total number of frames, that trial was excluded from subsequent analyses. Note that this comparison was conducted only on the frames of the subject phrase. A total of 57 trials (out of 2693, 2.1%) were removed due to the insufficient number of F0 frames in the subject phrase. These cases were predominantly

found in one specific participant (35 out of 57 trials), who produced creaky voice towards the end of the subject phrase.

For the rest of the trials, the F0 trajectories were smoothed, and the missing values were interpolated using a cubic spline method. See the green dashed lines in Figure 3.5. After smoothing and interpolation, F0 values were inspected again to make sure that there were no values that went above the gender-specific F0 range (M: 40-200 Hz, F: 100-320 Hz), which may have occurred due to extrapolation. There were three trials (out of 2693, 0.1%) that had F0 values outside this range, and these trials were excluded from the analyses.

3.2.5 Measurements

3.2.5.1 F0 measures in the subject phrase

An F0 contour of each NP in the subject phrase was linearly time-warped, and its average was plotted to examine the intonational pattern of each participant. This was to identify the common accentual pattern among participants, and the data of those who exhibited such pattern were subject to various analyses. Figure 3.6 presents the average time-warped F0 contours of each experimental condition in each participant. The F0 values in this figure were recentered within each participant using their global F0 mean, to facilitate across-participant comparison. Although the F0 contour was time-warped by phrase, it was connected for visualization; the conjunction *"and"* was included at the beginning of the second and third NPs.

By qualitatively inspecting Figure 3.6 together with audio, one major accentual pattern was identified; it was the pattern produced by the seven participants in the first

and second rows of Figure 3.6. In their productions, each NP had an F0 valley that occurred at the numeral, a peak at the color, and another valley at the animal. These three F0 landmarks were thus the target of the analyses.

There were other intonational patterns in the data, as shown in the final row of Figure 3.6. PA08 showed a rising intonation at the end of each NP, and PA09 placed the F0 peak at the numeral of the NP, instead of the color as in the major accentual pattern. The data from these participants could be analyzed, but since there was only a single participant for each distinct pattern, they were not included in the analyses. Unlike PA08 and PA09, PA10 did not exhibit a consistent F0 pattern throughout the experiment session, as they produced the rising intonation in some trials and the F0 peak in numeral in other trials (i.e. the mixture of intonational patterns found in PA08 and PA09). PA11 also did not show a consistent F0 pattern, and moreover, the F0 trajectories of this participant were not as dynamic as those of other participants. Due to the insufficient samples and inconsistent F0 patterns, the data from these four participants were not included in the analyses.

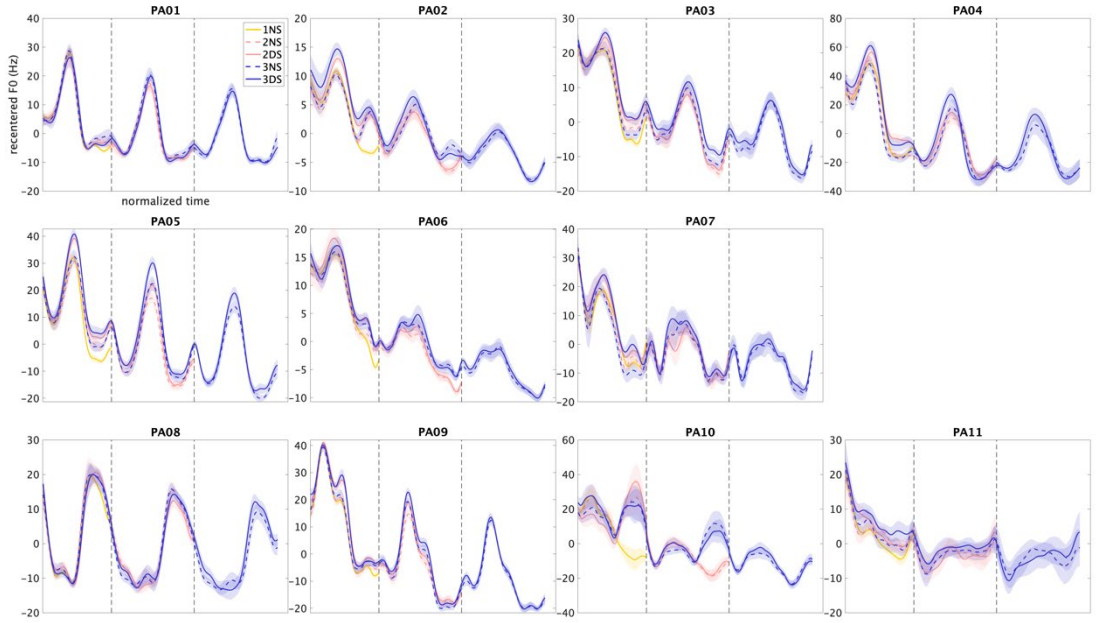

Figure 3.6. Smoothed/interpolated F0 contours that were time-warped by each subject NP. F0 (y-axis) was recentered within each participant using their global F0 mean. The vertical dashed lines mark the end of the time-warped NP; the conjunction "and" is included in the beginning of the second and third NP. The yellow lines show the F0 contours of 1NS, the pink lines show those of 2N/DS, and the blue lines indicate 3N/DS. In addition, the dashed lines show the delayed stimuli conditions (DS), while the solid lines show no-delayed stimuli conditions (NS). For each contour, the line in the middle shows the mean F0 values, and the shades indicate 95% confidence intervals.

The F0 landmarks (i.e. F0 valley preceding the peak, F0 peak, F0 valley following the peak) were identified as follows. For each NP, the highest peak was first identified. I then searched for the lowest valley in the region from the start of the NP to the peak, and another valley in the region from the peak to the end of the NP. If the peak was found at the edges of the NP – specifically, within the initial and final five frames of the given NP, the landmark measures were not recorded for that trial, as it suggests that the F0 contour did not conform to the major accentual pattern. For all other trials, F0 values of the landmarks as well as the values for the rises and falls (i.e. the F0 ranges between the peaks and valleys) were recorded. Figure 3.7 presents the sample F0 contour of NP1,

along with the markings of five F0 dependent variables: F0 peak, valley preceding the peak, valley following the peak, rise, and fall. The F0 landmarks are referred to as Vpre – P – Vpost and F0 rises and falls as R – F throughout this dissertation. The NP location could be added in the labels: for instance, the landmarks and rises/falls at NP1 are referred to as Vpre1 – P1 – Vpost1 – R1 – F1.

Among these five F0 dependent variables, the peak, valley preceding the peak, and fall (i.e. P, Vpre, F) are particularly important, as they were considered to indirectly represent H and L pitch targets and register span, respectively, in subsequent analyses. Note that the register span was associated with the F0 fall (i.e. F0 range between the peak and the *following* valley), rather than the rise (i.e. F0 range between the peak and the *preceding* valley). This was because the F0 value of the Vpost was in general lower than the value of the Vpre (see Figure 3.6), and thus, the floor of the register could be better characterized with the Vpost measure.

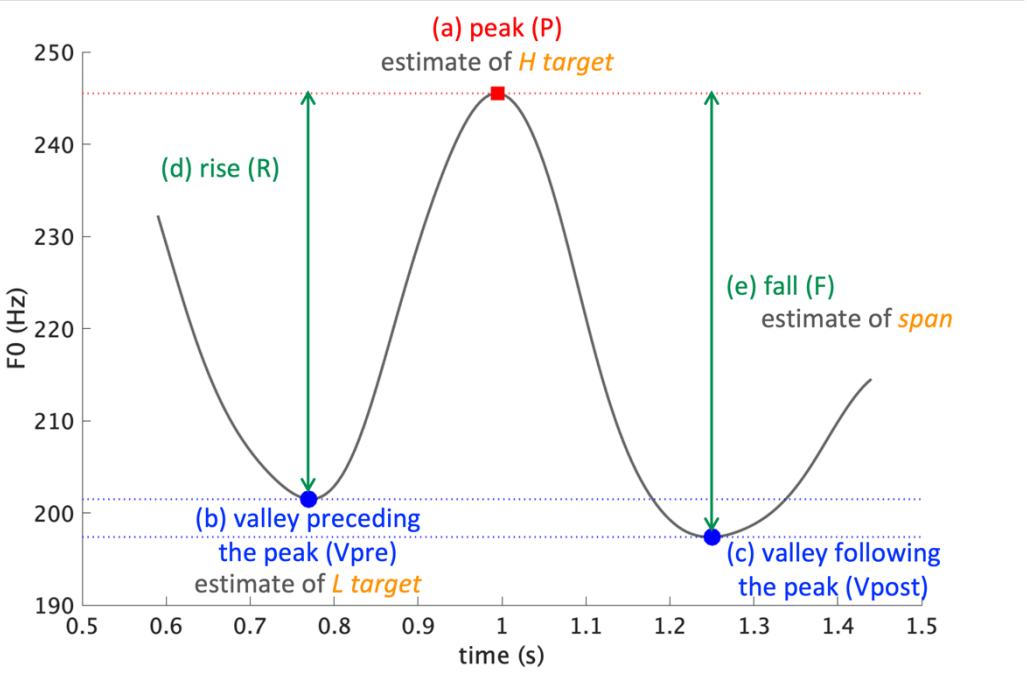

Figure 3.7. F0 dependent variables examined in the study. An example of a smoothed and interpolated F0 contour of NP1 is shown (gray line). (a): F0 peak, (b)/(c): F0 valley preceding/following the peak, (d): F0 rise, (e): F0 fall. The abbreviations used for each measure are presented in parentheses. In subsequent analyses, (a), (b), and (e) are considered as the estimates of H and L pitch targets and register span, respectively.

The segmental anchor for each of the F0 landmarks (i.e. Vpre, P, Vpost) was identified by examining the difference between the timepoints of the landmarks and the vowel onsets of the numeral, color, and animal. Figure 3.8 shows the distribution of such differences. In general, the difference between the vowel onsets and the timepoints of the landmarks was smallest in the numeral for Vpre, the color for P, and the second vowel of the animal for Vpost; in Figure 3.8, the medians of the boxplots of the numeral V, color V, and animal V2 were closest to 0 in Vpre, P, and Vpost, respectively. This result suggests that the Vpre, P, and Vpost were aligned to the vowel onsets of the numeral, color, and the second syllable of the animal. The formal notations for the F0 pattern of each NP would be L+H* which is followed by an L*, L-, or L%, depending

on the prosodic phrasing; these notations will be further discussed in Section 3.4.5. Note that the Vpre1 and P1 occurred a little later than their segmental anchors, which can be found in the first two panels of Figure 3.8.

Given these segmental anchors, outliers of F0 landmarks were identified using a mixed-effects linear regression. For each F0 landmark, a linear model was fit between the timepoints of the landmarks and the onsets of the segmental anchors with the participant as a random intercept. The datapoints whose absolute residuals were larger than 2.326 (the 1/99 percentiles of a normal distribution) were excluded from the analyses. For a given trial, if any of the F0 landmarks (i.e. Vpre/P/Vpost) were determined to be outliers, F0 rises and falls which were calculated based on those landmarks were also excluded from the analyses.

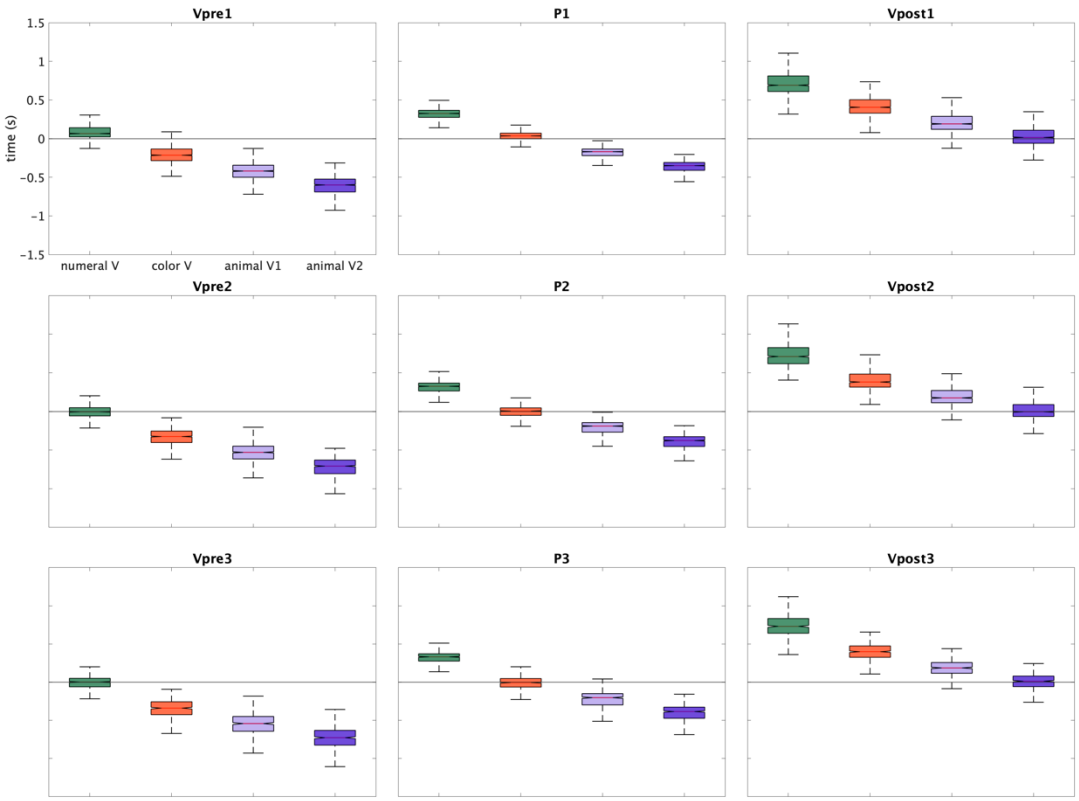

Figure 3.8. Distributions of differences between the timepoints of the landmarks and the vowel onsets. Each row shows the distributions of NP1, NP2, and NP3, and each column shows those of valleys preceding the peaks (Vpre), peaks (P), and valleys following the peaks (Vpost). In each panel, the green box shows the difference between the timepoints of the landmarks and the vowel onset of the numeral, the orange box shows the difference with the vowel onset of the color, and the blue boxes show the differences with the onsets of the first and second vowels of the animal. The horizontal line marks 0, which indicates that there is no temporal gap between the landmarks and the start of the vowel.

Together with the individual F0 measures of each NP, differences of measures across NPs were also examined. Specifically, differences in F0 peaks, F0 valleys preceding the peak, and F0 falls between NP1-NP2 and NP2-NP3 were recorded. As mentioned above, each of these F0 measures is considered to indirectly represent H and L pitch targets and register span; thus, it is expected that their differences across NPs (specifically, between NP1 and NP2) would demonstrate how participants control pitch

targets and register as they encounter delayed stimuli. Figure 3.9 represents the difference measures between NP1 and NP2.

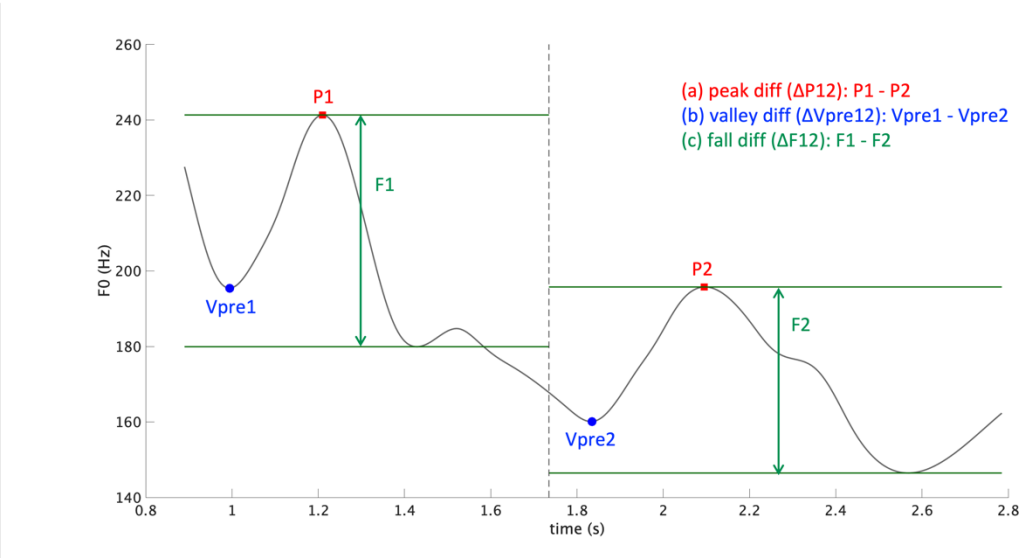

Figure 3.9. Differences of F0 measures between NPs examined in the study. An example of a smoothed and interpolated F0 contour of NP1 and NP2 is shown (gray line); the vertical dashed line marks the end of NP1. (a): difference between F0 peaks (P1-P2), (b): difference between F0 valleys preceding the peaks (Vpre1-Vpre2), (c): difference between F0 falls (F1-F2).

3.2.5.2 F0 measures in the verb phrase

Although the main targets of the analyses were the F0 landmarks and rises/falls in the subject phrase, I additionally measured the lowest F0 value of the VP (VPmin) as well as the maximum F0 value preceding that minimum (VP max) to characterize the effects of sentence length and delayed stimuli presentation on the utterance-final F0. Since F0 values were missing in a lot of the frames towards the end of the utterance, presumably due to the irregular and creaky phonation that is observed utterance-finally, it was considered more appropriate to examine the maximum and minimum F0 values rather than the whole F0 contour itself.

The identical procedure of identifying segmental anchors was applied to the two F0 values associated with the VP. Specifically, the timepoints of the VPmax and VPmin were compared with the vowel onsets of the words *"live", "in", "the",* and *"zoo*". The smallest difference for VPmax was found in the vowel onset of *"live",* and it was the vowel onset of *"the"* for VPmin. A mixed-effects linear regression was conducted as mentioned above (i.e. DV: timepoints of VPmax or VPmin; IV: onsets of segmental anchors, with the random intercept of participants), and the same threshold was used to identify outliers of VPmax and VPmin, which were excluded from the analyses.

3.2.5.3 Duration measures

Along with the F0 measurements, the durations of each subject NP and words within the NPs were also measured. The durations were measured from the data of 11 participants, regardless of their accentual patterns. At a phrase-level, the durations of NPs and the between-NP intervals – i.e. the duration of *"and"* and the silence preceding and following it (if present) – were measured. At a word-level, the durations of each word – i.e. numeral, color, animal, *"and"* – were measured. The durations of words and phrases were obtained from acoustic segmentation; see Figure 3.10 for an example. No additional outlier removal process was carried out, as the trials with extreme word and silence interval durations were already removed for potential disfluencies (cf. Section 3.2.3).

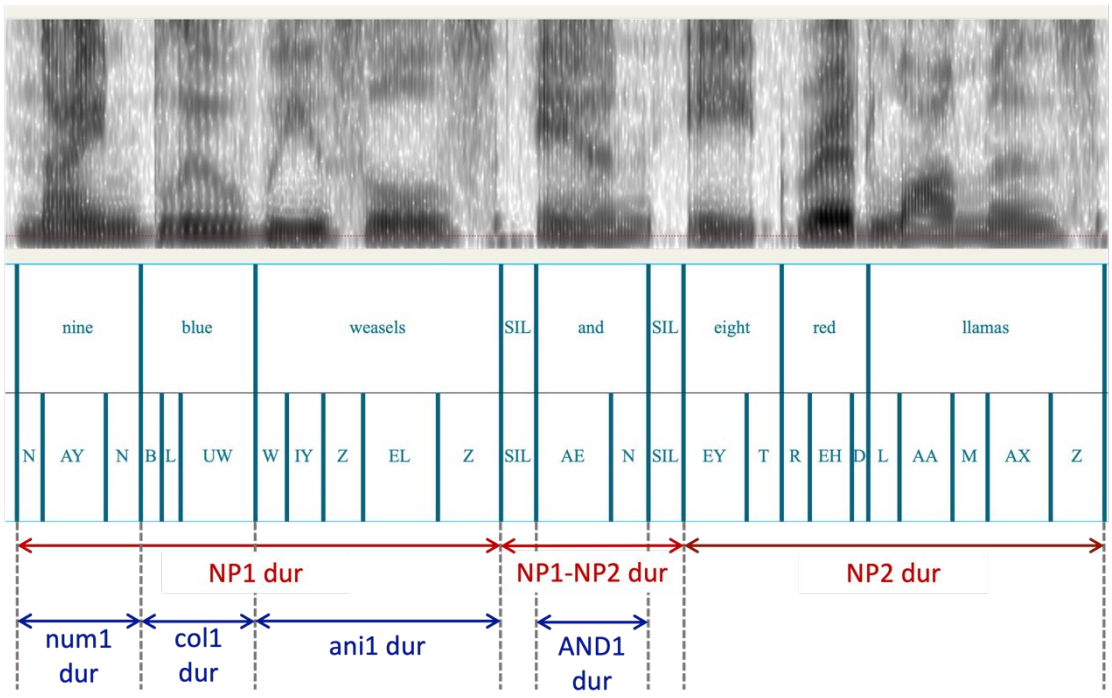

Figure 3.10. An example of the forced alignment. For each trial, the durations of NPs (NP1 dur and NP2dur in this example), between-phrase intervals (NP1-NP2 dur), words within NPs (num1 dur, col1 dur, ani1 dur), and conjunction (AND1 dur) were measured.

3.2.5.4 Summary

Overall, Table 3.5 summarizes all dependent variables examined in this study. The first set of F0 variables in the table were measured at each subject NP. Since an F0 contour may have multiple peaks and valleys within NP, F0 peaks and valleys examined in the analyses indeed are the highest F0 peak and lowest F0 valleys found in the data. The differences in the F0 variables across NPs were measured only for the F0 peaks, preceding F0 valleys, and falls as mentioned above, as they are the best estimates of H and L pitch targets and register span.

Table 3.5. Summary of dependent variables examined in the study. The top table lists F0 measures, and the bottom table lists duration measures. The F0 variables of the subject phrase and word durations were measured at each NP, and the F0 differences across NPs were measured between NP1-NP2 and NP2-NP3; the abbreviation of these variables may contain information about NP location, and the example for the measures of NP1/NP1-NP2 is presented in the table.

name	abbreviation	description
F0		
subject phrase (for each NP)		
F0 peak	P (e.g. P1)	the (highest) F0 peak within NP
F0 valley preceding the peak	Vpre (e.g. Vpre1)	the (lowest) F0 valley preceding the peak
F0 valley following the peak	Vpost (e.g. Vpost1)	the (lowest) F0 valley following the peak
F0 rise	R (e.g. R1)	the F0 range between the peak and the preceding valley
F0 fall	F (e.g. F1)	the F0 range between the peak and the following valley
subject phrase (differences across NPs)		
F0 peak difference	ΔP (e.g. ΔP12)	the difference of F0 peaks between NPs
F0 valley difference	ΔVpre (e.g. ΔVpre12)	the difference of F0 valleys preceding the peaks between NPs
F0 fall difference	ΔF (e.g. ΔF12)	the difference of F0 falls (peak – following valley) between NPs
verb phrase		
F0 maximum of VP	VPmax	the F0 maxima of VP
F0 minimum of VP	VPmin	the F0 minima of VP

duration		
phrase		
phrase duration	NP1 dur, NP2 dur, NP3 dur	the duration of NP
between-phrase interval duration	NP1-NP2 dur, NP2-NP3 dur	the duration of the interval between NPs, which include the conjunction *"and"* and silence before/after it (if present)
word (for each NP)		
duration of numeral	num dur (e.g. num1 dur)	the duration of numeral within NP
duration of color	col dur (e.g. col1 dur)	the duration of color within NP
duration of animal	ani dur (e.g. ani1 dur)	the duration of animal within NP
duration of *"and"*	AND1 dur, AND2 dur	the duration of conjunction "*and"*

3.2.6 Data analysis

3.2.6.1 Statistical analysis

To test the effects of sentence length and delayed stimuli presentation on our dependent variables, several mixed-effects linear regression models were fit to the data. The model formula/terms differed by where in the utterance the measurements were taken. See Table 3.6 for the full detail of statistical models.

The "Group I. NP2, NP1-NP2 measures" shows the statistical model tested on the measurements from NP2 or the interval between NP1 and NP2. The relevant trials were those in conditions 2DS, 2NS, 3DS, and 3NS. For each dependent variable, a linear mixed-effects model was fit with the fixed effects of sentence lengths (1/2/3 NP) and stimuli delay (NS/DS), their interactions, and the random intercepts of participants. This maximal model was subsequently compared to the model that lacked the interaction term and then with the models that lacked either fixed effect of length or stimuli delay. If the maximal model (with the interaction term) was found to be significant, further analyses on the significance of fixed effects were not conducted. The model comparison aimed to identify the significant terms in the model, and it was conducted through a loglikelihood test. Note that for F0 variables, statistical tests were conducted on the original F0 values; yet, for the graphical representation of the results in the next section, F0 values that were recentered using the participant's global F0 mean were used.

For the F0 or duration variables in NP3 and the interval between NP2 and NP3 – Group II in Table 3.6, only the delay effect could be tested. A linear regression model with the fixed effect of delayed stimuli presentation (DS/NS) and the random intercepts of participants was fit to the measures taken from this region.

The rest of the table presents statistical models that were fit to the measures taken from NP1. For the early NP1 F0 measures – namely, F0 values of Vpre1, P1, and R1, the effect of sentence length was tested with an alternative coding for the experimental conditions (Group III). Specifically, the experimental conditions were coded according to the number of stimuli presented at the beginning of the trial – thus, condition 1NS, 2DS, 2NS, 3DS, and 3NS were coded as 1Pi (one initial stimulus), 1Pi, 2Pi (two initial stimuli), 1Pi, and 3Pi (three initial stimuli), respectively. This was based on the empirical observation (Figure 3.11 in Section 3.3.1.1 below), in which the measures of 1NS, 2DS, and 3DS conditions did not vary significantly from each other. In principle, it is not impossible for early NP1 measures to be affected by the appearance of delayed stimuli, but I could not find any evidence for it in the data. Note, however, that the duration measures of NP1 which were aligned to the F0 measures of Vpre1 and P1 (i.e. num1 dur, col1 dur) were not included in this group, but rather, they were tested with the statistical models in Group IV. This is because given that the delayed stimuli were presented on average 86.4 ms after utterance initiation, it is possible that the effect of delayed stimuli may have already showed up in these duration measures.

For the rest of the F0 measures of NP1 (i.e. F0 values of V1post, and F1), the duration measures associated with NP1 (i.e. NP1 dur, num1 dur, col1 dur, ani1 dur), and the F0 measures at VP, statistical analyses were conducted on different subsets of the data (Group IV). The main reason for dividing the data into subsets was that the experiment lacked 1DS condition. For the measures from the trials in 2DS, 2NS, 3DS, and 3NS conditions, the model specified in IV-(i) in Table 3.6 was fit to the data. As in the analysis in the Group I, the maximal model was compared with the models that

lacked an interaction effect or fixed effect through a loglikelihood test. On the other hand, for the subset of the data which was composed of 1NS, 2NS, and 3NS trials (IV-(ii)), the effect of sentence length was examined with the random intercepts of participants. Lastly, the effect of utterance length was tested for the measures of 1NS, 2DS, and 3DS trials (IV-(iii)).

Table 3.6. Summary of statistical models. For each group, the measurements that were subject to a given statistical test are listed along with the model formula and the explanation on model terms. DV in the formula indicates the dependent variable.

I. NP2, NP1-NP2 measures	
measures	F0: P2, Vpre2, Vpost2, R2, F2, ΔP12, ΔVpre12, ΔF12 Dur: NP2 dur, NP1-NP2 dur, num2 dur, col2 dur, ani2 dur, AND1 dur
model formula	DV ~ 1 + length*delay + (1|part)
model terms	- interaction effect between length and delay - fixed effect of length - fixed effect of delay - random intercepts of participants
II. NP3, NP2-NP3 measures	
measures	F0: P3, Vpre3, Vpost3, R3, F3, ΔP23, ΔVpre23, ΔF23 Dur: NP3 dur, NP2-NP3 dur, num3 dur, col3 dur, ani3 dur, AND2 dur
model formula	DV ~ 1 + delay + (1|part)
model terms	- fixed effect of delay - random intercepts of participants
III. NP1 measures (all trials)	
measures	F0: P1, Vpre1, R1 cf. alternative coding: 1Pi, 1Pi, 2Pi, 1Pi, 3Pi (1NS, 2DS, 2NS, 3DS, 3NS)
model formula	DV ~ 1 + length + (1|part)
model terms	- fixed effect of length - random intercepts of participants
IV. NP1 measures (trials in selected conditions)	
measures	F0: Vpost1, F1, VPmax, VPmin Dur: NP1 dur, num1 dur, col1 dur, ani1 dur
(i) trials in 2DS, 2NS, 3DS, 3NS conditions	
model formula	DV ~ 1 + length*delay + (1|part)
model terms	- interaction effect between length and delay - fixed effect of length - fixed effect of delay - random intercepts of participants
(ii) trials in 1NS, 2NS, 3NS conditions	
model formula	DV ~ 1 + length + (1|part)
model terms	- fixed effect of length - random intercepts of participants
(iii) trials in 1NS, 2DS, 3DS conditions	
model formula	DV ~ 1 + length + (1|part)
model terms	- fixed effect of length - random intercepts of participants

3.2.6.2 Analysis of F0 control

Additional analyses were conducted to investigate the speakers' F0 control mechanism. The analyses focused on the F0 measures associated with NP1 and NP2, as these are the critical regions that speakers would control their F0 according to the manipulations of sentence length and delayed stimuli presentation.

The first analysis was on the variance of F0 measures. The variances of F0 peaks, valleys, and ranges were calculated for each participant and experimental condition, and the sum of the variances of the peaks and valleys was compared with the variance of the ranges. Specifically, at each NP, (i) the sum of the variances of P and Vpre was compared with the variance of R (i.e. difference between P and Vpre), and (ii) the sum of the variances of P and Vpost was compared with the variance of F (i.e. difference between P and Vpost). This comparison was intended to find out whether peaks and valleys were independently controlled.

In the second analysis, the correlation between F0 peaks and valleys was examined at each NP. In particular, the correlation between F0 peaks and F0 valleys following the peaks, the ranges of which were considered to estimate the register span, was examined at each NP – i.e. the correlation between P1-Vpost1, P2-Vpost2. The correlations were calculated separately in DS and NS trials within each participant.

The last analysis examined which of the three F0 measures that are considered to represent H pitch targets, L targets, and register span best predicts the delayed vs. no-delayed experimental condition. As mentioned above, F0 peaks were considered as the estimates of H targets, F0 valleys preceding the peaks as the estimates of L targets, and F0 falls as the register span (Section 3.2.5.1), and thus, the analysis focused on these

measures. In particular, for each NP, three logistic regression models were fit to the data to predict DS vs. NS conditions, and the regression models were compared using the Akaike Information Criterion (AIC). For NP1, 1NS trials were excluded, as there was no 1DS counterpart. The models included each of the F0 measures (i.e. P, Vpre, F) and sentence length as fixed effects, their interactions, and the random interactions, slopes, and intercepts for participants.

3.3 Results

Two different sets of analyses were conducted on the experiment data: the first set aimed to examine whether speakers show evidence for the pre-planned and adaptive F0 control, and the second set aimed to identify the control parameter (*targets* vs. *register*) that speakers used to produce F0 variations.

In the first set of the analyses, F0 measures of each NP and their differences across NPs were examined. Note that these analyses were conducted on the participant-pooled data with a goal of identifying the common strategy of F0 control; for the individual differences among participants, see discussions in Section 3.4.3. First, analyses of the F0 landmarks (Vpre, P, Vpost) and the ranges between them (R, F) showed a significant effect of sentence length and delayed stimuli presentation. Specifically, participants produced a higher P1 and V1 and a wider R1 in sentences with more initial stimuli (i.e. 2Pi, 3Pi). This suggests that participants were sensitive to the initial sentence length and pre-planned their F0 control according to that information. Second, regarding the effects of stimulus delay, the difference between the F0 peaks across NPs (i.e. ΔP12, ΔP23)

was smaller in the condition in which the stimuli were delayed. In particular, F0 values of the peaks in general decreased from NP1 to NP2 and NP2 to NP3, but the amount of decrease was larger in NS trials than DS trials. The results provide evidence that participants were also sensitive to the changes in the length that were made after utterance initiation and adjusted their F0 control accordingly.

In the second set of the analyses, which assessed the F0 control hypotheses (*target* vs. *register*), analyses mostly pointed to the *register*-control, although the results were inconsistent in the comparison of condition-prediction models. In the (i) variance analysis, the sum of the variances of F0 peaks and valleys was larger than the variance of ranges in the majority of participants and conditions, which provided support for the *register*-control hypothesis. In the (ii) correlation analysis, a positive correlation was observed between F0 peaks and valleys, and a moderate to high correlation was found for many participants/conditions, which provided evidence for the *register*-control hypothesis. The results from the (iii) model comparison were yet difficult to interpret, as they showed evidence for *target*-control at NP1 and *register*-control at NP2.

I first present F0 variations that were induced by sentence length and stimulus delay in Section 3.3.1. I then introduce the results from the correlation and variance analyses and model comparisons in Section 3.3.2, all of which were designed to examine the F0 control hypotheses – i.e. *target* vs. *register*-control. Lastly, I present some other findings on the data, specifically regarding the F0 measures associated with VP (i.e. VPmin, VPmax) and phrase/word durations (Section 3.3.3). Most of the figures in this section have y-axis as F0 values that were recentered using the global F0 mean of each participant. However, statistical tests were conducted on the original F0 values with

random intercepts of participants, as stated in the previous section. The full list of regression model coefficients and statistical significance from the analyses of subject NP F0 measures (Section 3.3.1) and phrase/word durations (Section 3.3.3.2) is provided in the Appendix. In this section, I introduce the subset of the table that is directly relevant to the findings.

3.3.1 Effects of sentence length and delayed stimuli presentation

3.3.1.1 Initial sentence length

Analyses of the F0 values of Vpre1, P1, and R1 showed significant effects of initial sentence length. In particular, F0 values of these measures increased in sentences with more initial stimuli. This suggests that participants made an utterance plan before production that considered the number of stimuli presented at the beginning of the trial. For the visualization of this effect, see Figure 3.11 and Figure 3.12. Figure 3.11 shows the average time-warped F0 contours of the subject phrase (top) as well as the average F0 values of the landmarks (bottom) of the seven participants who showed similar accentual patterns; Figure 3.12 shows the distributions of Vpre1, P1, and R1 for a more detailed comparison between the conditions. From the figures, we could find that the trials in 1NS, 2DS, and 3DS conditions patterned together in all three F0 measures; this provided a basis of grouping these conditions together for the statistical analyses (i.e. 1Pi). The F0 measures of 1NS, 2DS, and 3DS trials, however, were different from those of 2NS and 3NS trials in that the values of the former were lower than the latter.

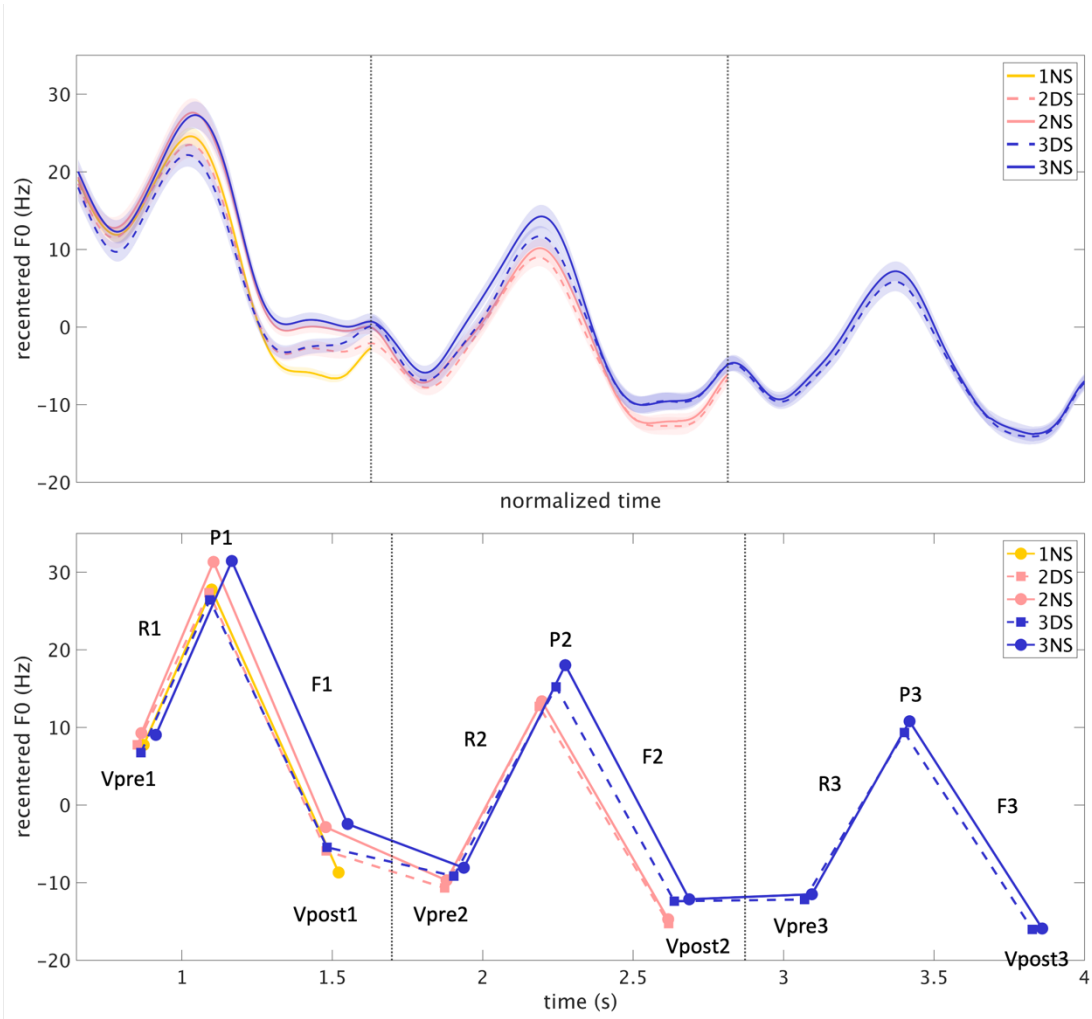

Figure 3.11. Average time-warped F0 contours (top) and F0 landmarks (bottom). The yellow line represents sentences with a single NP (1NS), the pink lines show those with two NPs (2NS/2DS), and the blue lines show those with three NPs (3NS/3DS). The dashed lines/square markers indicate DS conditions, and the solid lines/circle markers indicate NS conditions. The vertical dotted lines represent NP boundaries; the lines in the top panel show the length of the time-warped NPs, and the lines in the bottom panel mark the timepoints in the middle of Vpost and Vpre that were averaged across conditions. F0 was recentered within each participant using their global F0 mean.

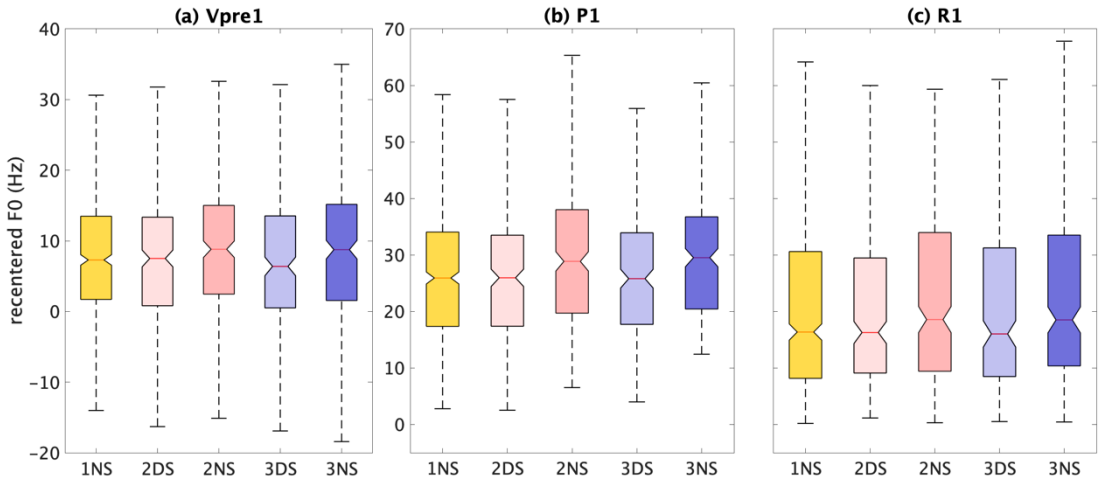

Figure 3.12. Distributions of Vpre1, P1, and R1 by experimental condition. The yellow box represents 1NS, the pink boxes represent sentences with two NPs (2DS/2NS), and the blue boxes represent those with three NPs (3DS/3NS). The lighter colors show DS conditions. For statistical analyses, the conditions were recoded according to the number of initial stimuli (i.e. 1Pi, 2Pi, 3Pi).

For the statistical analyses of these F0 variables, the experimental conditions were recoded according to the number of initial stimuli – i.e. 1Pi, 1Pi, 2Pi, 1Pi, 3Pi instead of 1NS, 2DS, 2NS, 3DS, 3NS (see Section 3.2.6.1). Table 3.7 presents the regression coefficients and their statistical significance. The result shows that the participants started with a higher F0 valley and F0 peak as well as a wider F0 range, when they were presented with more initial stimuli. Additional models were fit to the trials with just two and three initial stimuli (2Pi, 3Pi), and a significant length effect was observed at P1; the coefficient of the condition 3Pi was 1.31 Hz ($p < 0.01$). This result further suggests that the participants distinguished two vs. three initial NPs when they were producing F0 peaks.

*Table 3.7. Regression model coefficients of Vpre1, P1, and R1. The three conditions – 1Pi, 2Pi, and 3Pi – were treated as a categorical variable, and the reference group was 1Pi (i.e. 1NS, 2DS, 3DS). The coefficients are in the unit of Hz. ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$.*

	Vpre1	**P1**	**R1**
2Pi (=2NS)	1.87**	3.94***	2.11**
3Pi (=3NS)	2.07**	5.06***	2.64**

F0 measures of Vpost1 and F1 also showed the effects of sentence length as well as delayed stimuli presentation. This result, however, was relevant to the markings of the end of the subject phrase, rather than the speakers' pre-planned F0 control. Thus, the analyses of these measures, along with the analyses of Vpost2 and F2, are presented in a separate section (Section 3.3.1.3). Also, for the inter-participant differences on how they varied F0 parameters according to the initial sentence length (as well as the changes in the length, which is introduced in the section below), see Section 3.4.3.

3.3.1.2 Delayed stimuli presentation

To examine how participants controlled F0 variables as they saw delayed stimuli, the differences of F0 peaks, valleys preceding the peaks, and falls across NPs were investigated. Among these measures, F0 peak difference between NP1 and NP2 (ΔP12) and between NP2 and NP3 (ΔP23) showed a significant effect of delay. Figure 3.13 shows the distribution of F0 differences by experimental condition. As can be seen from the panels (a) and (b), F0 peaks in general decreased from NP1 to NP2 and from NP2 NP3: the medians of the boxplots in panels (a) and (b) were all above 0. Yet, the amount of decrease differed by condition such that the decrease was larger in the NS conditions compared to the DS conditions: the medians of the darker pink and blue boxes were higher than those of the lighter boxes. This was also confirmed in the statistical analysis,

as the coefficient of the NS conditions (compared to the DS conditions) was 2.78 Hz ($p < 0.001$) for ΔP12 and 1.49 Hz ($p < 0.05$) for ΔP23. The significant effect of delayed stimuli presentation was, however, not found in other F0 measures – i.e. ΔVpre12, ΔVPre23, ΔF12, and ΔF23.

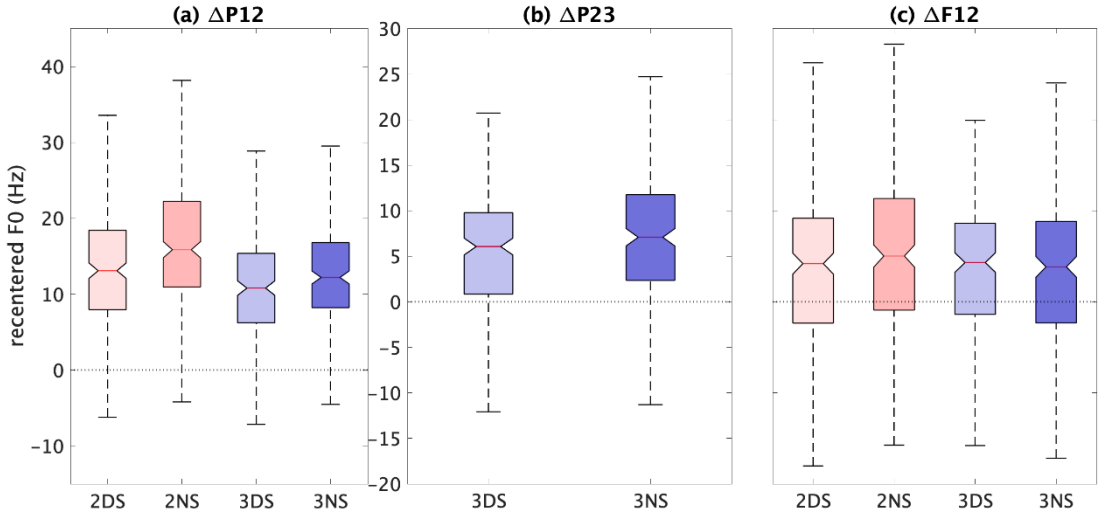

Figure 3.13. Distributions of ΔP12, ΔP23, and ΔF12 by experimental condition. The horizontal dotted line marks 0, which shows that the F0 values were identical across phrases.

F0 variations across NPs not only differed by the occurrence of delayed stimuli, but also by sentence length; the effect of length was particularly observed in ΔP12 and ΔF12. See panels (a) and (c) in Figure 3.13. In (a) ΔP12, F0 difference between NPs was smaller in longer sentences: the medians of the blue boxplots were lower than their pink counterparts. A similar pattern was observed in (c) ΔF12, though the difference was rather subtle, and there were a number of cases where the fall increased in NP2 compared to NP1 (ΔF12 < 0). Statistically, a significant effect of length was observed in both cases: the coefficient of the trials with three stimuli (compared to the trials with two stimuli) was -3.46 Hz ($p < 0.001$) in ΔP12 and -1.83 Hz ($p < 0.01$) in ΔF12.

Overall, the findings can be summarized as follows: participants in general lowered the F0 peak or compressed the F0 range across NPs. The important finding is that they adjusted the amount of decrease or compression according to the experimental condition such that they lowered F0 peaks or compressed F0 ranges to a *lesser* extent when they encountered delayed stimuli and/or when they had to produce longer sentences.

The careful adjustment of F0 peaks and ranges across NPs also had an influence on the F0 measures of NP2 and NP3. We have observed in Section 3.3.1.1 that, at NP1, the F0 measures of trials with a single NP stimulus (i.e. 1Pi – 1NS, 2DS, 3DS) differed significantly from the trials that had two or three initial stimuli (i.e. 2Pi – 2NS, 3Pi – 3NS). However, at NP2 and NP3 – i.e. after the delayed stimuli were presented, the difference between the DS and NS conditions diminished, and they tended to pattern together. In Figure 3.11, the F0 trajectories of 2DS and 2NS conditions and their landmarks were almost similar at NP2, and those of 3DS and 3NS also became similar towards NP3. This is also evident in Figure 3.14, which presents the distributions of P2, P3, F2, and F3 by experimental condition: the distributions of F0 peaks and falls of 2DS and 2NS trials were almost similar in panel (a) and (c); the distributions of 3DS and 3NS trials showed a smaller difference towards NP2 to NP3 – i.e. from (a) to (b) and from (c) to (d).

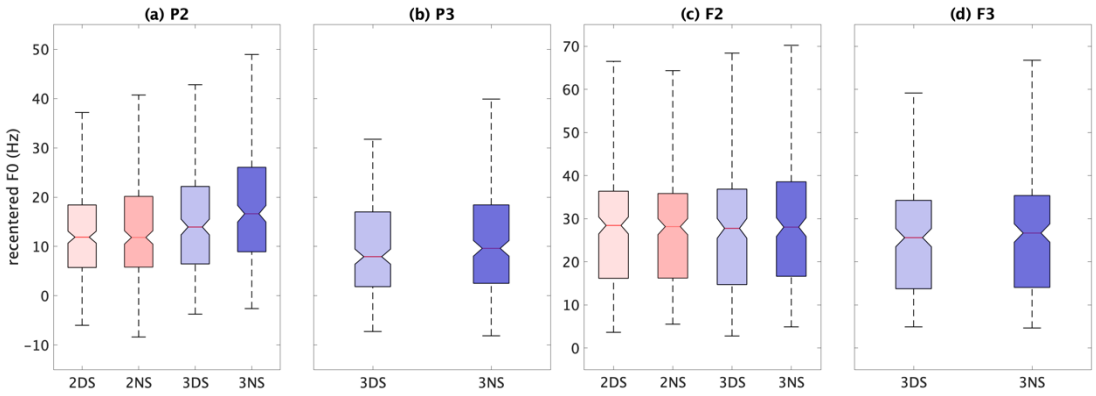

Figure 3.14. Distributions of P2, P3, F2, and F3 by experimental condition.

This result was also confirmed in the statistical analyses. Except for the Vpre2, which showed a significant effect of stimulus delay, P2 and R2 showed a significant interaction effect between sentence length and stimulus delay. At NP3, the effect of delayed stimuli presentation was found in the Vpre3, P3, and Vpost3 measures, although the coefficient difference between DS and NS conditions was around 1 Hz.

These observations suggest that the participants had pre-planned DS and NS trials differently before production (which was manifested in the F0 measures at NP1); yet, once they saw the delayed stimuli, they adapted to the changes, such that the production of the two conditions became similar over the course of the utterance. I argue that the adjustments of F0 peaks and ranges across NPs which differed by delayed stimuli presentation resulted in a similar pattern of DS and NS trials in NP2 and NP3.

3.3.1.3 NP-final F0 measures

Analyses of the NP-final F0 variables – i.e. Vpost and F – showed the effects of utterance length and delayed stimuli presentation, yet in a way different from the other F0 measures. While the F0 values of Vpre1, R1, and P1 did not exhibit a significant difference among 1NS, 2DS, and 3DS conditions (cf. Figure 3.11, Figure 3.12), which

was why they were all coded identically as 1Pi in the statistical analyses, the three conditions showed a crucial difference in the Vpost1 and F1 measures. In particular, F0 values of Vpost1 were lower, and the values of F1 were larger in 1NS trials compared to 2DS and 3DS trials; the Vpost1 measures of 1NS trials were also lower than those of 2NS and 3NS trials. This result is confirmed in the first column of Table 3.8 and panel (a) in Figure 3.15. In sum, the Vpost1 measure was the lowest in the stimuli with a single NP (1NS), which was followed by 2DS and 3DS, and then 2NS and 3NS.

Similarly, the Vpost measure at NP2 was lower in the stimuli that had two NPs compared to those with three NPs. In panel (b) in Figure 3.15, the medians of the 2DS and 2NS trials are lower than the 3DS and 3NS trials. A statistically significant effect of length was also observed in Vpost2, where the coefficient difference between sentences with 2NPs and 3NPs was 1.57 Hz ($p < 0.001$). The F2 measures showed an interaction effect.

Together with the analyses of Vpre1 and F1, these results altogether suggest that the Vpost (as well as F) is relevant to the markings of the end of the subject phrase. In particular, participants lowered the final F0 valley of an NP much further, if the given NP was the final NP of the subject phrase. This resulted in a particularly low Vpost measure of 1NS trials in NP1 and 2NS/DS trials in NP2.

*Table 3.8. Regression model coefficients of Vpost1, F1, Vpost2, and F2. Both sentence length and delay variables were treated as categorical variables, and the reference group was the shortest length and DS condition. The coefficients are in the unit of Hz. In the table, - indicates that the effect was not significant. The grey cells show that the effects were not tested for the given variable. ***: p < 0.001, **: p < 0.01, *: p < 0.05.*

	Vpost1	F1	Vpost2	F2
(i) trials in 1NS, 2NS, 3NS conditions				
2NS	5.79***	-1.34**		
3NS	5.65***	-		
(ii) trials in 1NS, 2DS, 3DS conditions				
2DS	2.51***	-2.48***		
3DS	2.13***	-1.73***		
(iii) trials in 2DS, 2NS, 3DS, 3NS conditions				
interaction	-	-	-	2.16*
length (3NPs)	-	0.98*	1.57***	
delay (NS)	3.29***	1.19**	0.67*	

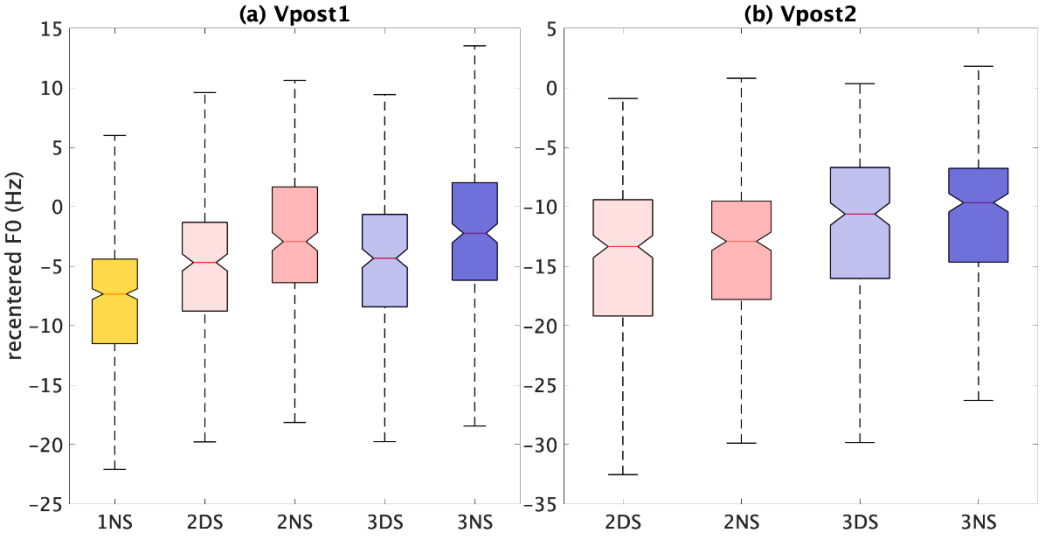

Figure 3.15. Distributions of Vpost1 and Vpost2 by experimental condition.

3.3.2 Investigation of F0 control hypotheses

3.3.2.1 Variance of F0 measures

Analyses of the variance of peaks/valleys/ranges showed that the F0 peaks and

valleys are not independent from each other, supporting the *register*-control hypothesis. For each NP, (i) the sum of the variances of F0 peaks and preceding valleys was compared with the variance of F0 rises, and (ii) the sum of the variances of peaks and following valleys was compared with the variance of F0 falls. It was hypothesized that if the sum of the variances of two measures (peaks/valleys) is substantially larger than the variance of the measure calculated from the two (ranges), it suggests that the two measures of interest (peaks/valleys) are influenced by a common mechanism, for which the register is an obvious candidate. Figure 3.16 visually presents the comparison results, by plotting the variance ratio: σ^2(range) / (σ^2(peak) + σ^2(valley)). The horizontal dashed line at 1 shows the case where the variance of the range is identical to the sum of the variances of peaks and valleys.

It was found that in both rises and falls and at both NPs, the variance of the range was smaller than the sum of the variances of peaks and valleys – i.e. the values were below 1 – in the majority of participants and conditions. In terms of the size of their differences, see the textbox in the bottom left corner of each panel, which shows the distribution of the variance ratio. If the variance ratio is closer to 0 (i.e. the variance of the range is very small compared to the sum of the variances), it suggests that the peaks and valleys are governed by a common mechanism; if the ratio, however, is closer to 1 (the variance of the range is similar to the sum of the variances), it means that the two measures are likely to be more independent. In all panels, the $0.5 \leq x < 0.75$ group had the greatest number of data points, which suggests that the variance of the range was about 50 to 75% of the sum of the variances of the peaks and valleys. This is a fairly large difference, which provides strong evidence for the *register*-control hypothesis.

An additional finding that can be seen in the figure is that there were some cases in which the variance ratio was quite large, especially in the comparison of P+Vpre vs. R at NP1. Around 34% of the data were in the $0.75 \leq x < 1$ group, which means that the range variance was at 75-100% of the sum of variances of peaks and valleys (i.e. smaller difference between σ^2(peak) + σ^2(valley) and σ^2(range)). This point will be further discussed in Section 3.4.4.

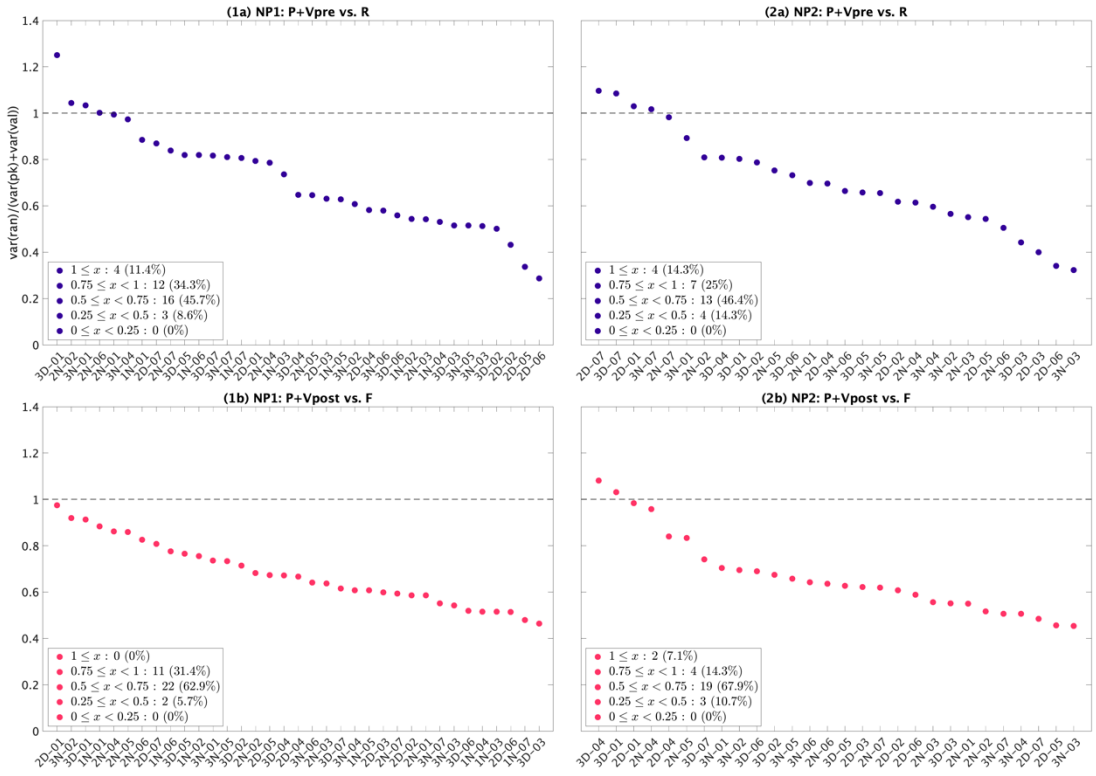

Figure 3.16. Comparison of the variance (var) of F0 measures. The figure plots variance ratio, which was calculated as the variance of the ranges divided by the sum of the variances of the peaks and valleys, for each participant and condition. The dots are sorted from the largest to the smallest ratio, and their labels indicate experimental condition – participant (e.g. 3D-01: 3DS trials in PA01). (a) var of peaks and preceding valleys vs. rises; (b): var of peaks and following valleys vs. falls. The left column shows the comparison at NP1, and the right shows NP2. The horizontal dashed line at 1 indicates cases where the var of the sum (peaks, valleys) and the var of the ranges are identical. The textbox in the bottom left corner shows the number of data points and percentage of each group.

3.3.2.2 Correlation between F0 measures

A positive correlation was found between the F0 peaks and the valleys following the peaks (i.e. P-Vpost) within each NP. When the adjusted R-squared values were examined – see Table 3.9, a moderate to high correlation was found in some participants and experimental conditions. There were indeed many cases where the R-squared values were above 0.2 (roughly, r = 0.45) – see the numbers in bold in Table 3.9. Given that the F0 range between the peaks and the following valleys was considered as an indirect estimate of register span, this result suggests that some participants in some contexts controlled register span to produce variations in F0. (cf. An alternative interpretation, in which participants controlled both H and L pitch targets is plausible but is not adopted; further discussions on this interpretation are provided in Section 3.4.4).

Table 3.9. Correlation between the F0 peaks (P) and the valleys following the peaks (Vpost), the range of which represents the register span, within each NP. The numbers in the table show the adjusted R-squared values between the two measures for a given participant and condition. The R-squared values larger than 0.20 are in bold.

	PA01	**PA02**	**PA03**	**PA04**	**PA05**	**PA06**	**PA07**
NP1							
NS	0.05	0.18	**0.26**	**0.30**	**0.27**	0.07	**0.25**
DS	-0.01	0.12	**0.23**	0.12	0.08	**0.24**	0.08
NP2							
NS	**0.21**	**0.26**	**0.38**	-0.01	**0.20**	**0.26**	**0.27**
DS	-0.01	**0.20**	**0.20**	-0.02	**0.33**	**0.25**	**0.21**

3.3.2.3 Model comparisons

Unlike the variance and correlation analyses, which provided strong support for the *register*-control hypothesis, the model comparison results were inconsistent and difficult to interpret. In NP1, the lowest AIC was found in the model which had F0 peaks

as the predictor, while in NP2, the lowest AIC was found in the model with F0 ranges as the predictor; the AIC values of each model are presented in Table 3.10.

This result seems to suggest that the participants controlled H pitch targets at NP1 but register span at NP2. However, we do not have any reason to believe that the participants vary F0 control parameters across phrases. Furthermore, it is possible that the effects of delayed stimuli presentation may differ between P/Vpre (early F0 measures) and Vpost (late F0 measure, which is related to F) especially at NP1, as participants could have started incorporating the delayed stimuli into the ongoing utterance at or before Vpost (but not at P/Vpre). This makes the results of the model comparison at NP1 less persuasive. These points would be further discussed in Section 3.4.4.

Table 3.10. Akaike Information Criterion (AIC) of the regression models that had either F0 peaks (P), valleys preceding the peaks (Vpre), or falls (F) as a predictor. The model with the smallest AIC at each NP is colored in yellow.

	P as predictor	**Vpre**	**F**
NP1	1404.48	1474.23	1472.65
NP2	1445.63	1439.50	1437.75

3.3.3 Other acoustic measures

3.3.3.1 F0 measures associated with VP

Analyses of the F0 maxima and minima of VP found a significant effect of sentence length. Figure 3.17 plots the average VPmax and VPmin calculated across participants, and Table 3.11 presents the statistical results. The general pattern was that the F0 measures associated with VP were higher in shorter sentences. F0 values of

VPmax were highest in 1NS trials, and the trials with two NPs (2DS/NS) exhibited higher F0 values than the trials with three NPs (3DS/NS). A small effect of delayed stimuli presentation was also observed, with higher VPmax values in NS trials. F0 values of VPmin also showed a significant effect of length. The 1NS trials again had the highest F0 values, and the difference between the trials with two NPs and three NPs was also observed. This result suggests that the utterance-final F0 values vary by the length of the sentence, contrary to the common finding that the speakers end an utterance with a similar, stable F0 regardless of the sentence length (e.g. Liberman & Pierrehumbert, 1984); see Section 3.4.5 for further discussion.

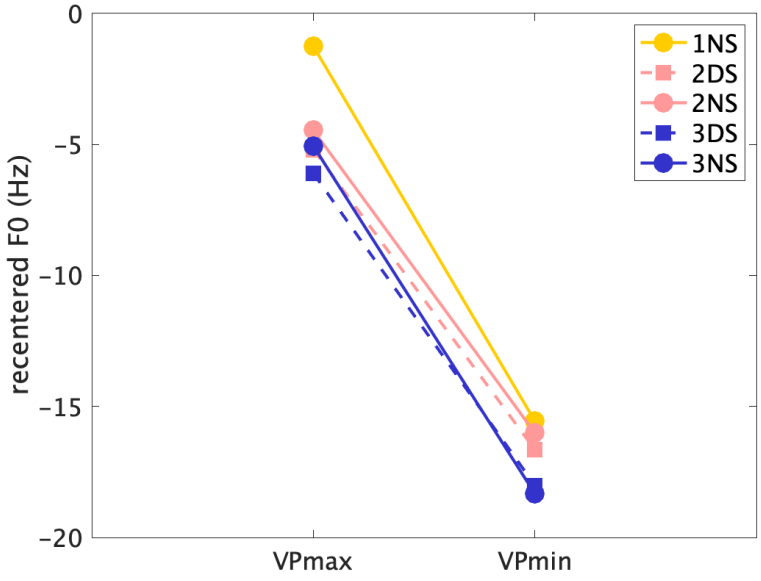

Figure 3.17. Average F0 maxima and minima of the verb phrase.

Table 3.11. Regression model coefficients of VPmax and VPmin. The coefficients are in the unit of Hz.

	VPmax	VPmin
(i) trials in 1NS, 2NS, 3NS conditions		
2NS	-3.59***	-
3NS	-4.59***	-3.6***
(ii) trials in 1NS, 2DS, 3DS conditions		
2DS	-4.23***	-1.37*
3DS	-5.52***	-3.45***
(iii) trials in 2DS, 2NS, 3DS, 3NS conditions		
interaction	-	-
length (3NPs)	-1.37***	-2.58***
delay (NS)	0.71*	-

3.3.3.2 Phrase and word durations

Analyses of phrase durations showed a significant effect of length in NP1, NP1-NP2 interval, and NP2. In particular, durations of these phrases/intervals were longer in sentences with more subject NPs – i.e. the NP1, NP1-NP2 interval, and NP2 durations were longer in sentences with three subject NPs compared to two NPs and compared to a single NP (in case of NP1 dur). Figure 3.18 shows the average durations of NPs and between-NP intervals calculated over 11 participants, with markings of locations where significant effects of length and stimulus delay were observed. Figure 3.19 provides the distributions of duration measures in those locations. In all panels of Figure 3.19, the medians of the stimuli with more subject NPs were higher than those with fewer NPs. The significant length effect is also confirmed in Table 3.12, which presents statistical results.

Regarding the delay effect, only the NP1-NP2 dur showed a significant difference by delayed stimuli presentation. Specifically, the duration was longer in conditions with

delayed stimuli at about 6.86 ms. This suggests that participants incorporated the delayed stimuli into their utterance mainly at this location, although the effect appears to be quite small. Nonetheless, together with a significant length effect found at this location, we might infer that it took longer for participants to incorporate two delayed stimuli (3DS) into their ongoing speech compared to a single delayed stimulus (2DS). The interval between NP2 and NP3, however, did not show a significant effect of delay. Presumably, this is because participants had a plenty of time to incorporate the delayed NP3 before reaching this region, as the delayed phrases were presented immediately after the utterance was initiated.

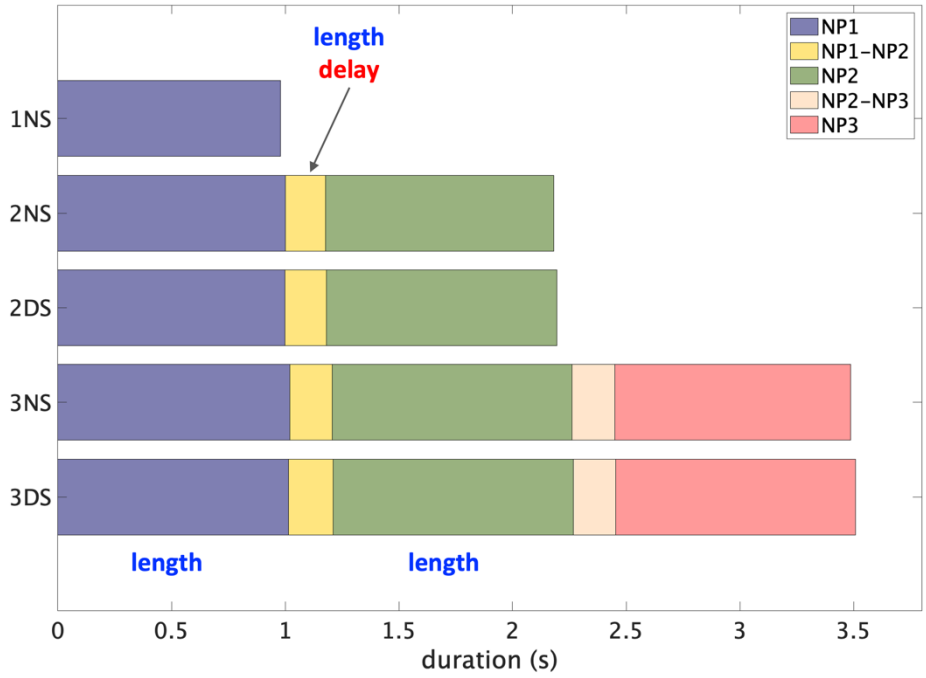

Figure 3.18. Mean durations of subject NPs and the intervals between NPs, with markings of locations that showed significant effects of length and/or delay. The blue, green, and pink bars show the average durations of NP1, NP2, and NP3, respectively. The yellow and light pink bars show the average interval durations between NP1 and NP2 and between NP2 and NP3, respectively. A significant effect of length was observed in NP1, NP1-NP2, and NP2 durs, and the effect of delay was found in NP1-NP2 dur.

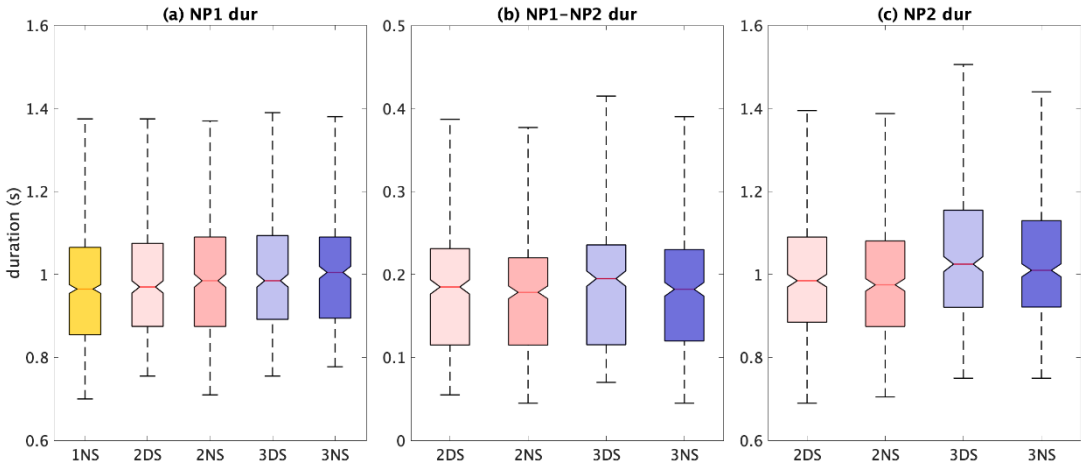

Figure 3.19. Distributions of NP1dur, NP1-NP2 dur, and NP2 dur, which showed a significant effect of length and/or delay, by experimental condition.

Table 3.12. Regression model coefficients of phrase durations. The coefficients are in the unit of ms.

	NP1 dur	**NP1-NP2 dur**	**NP2 dur**
(i) trials in 1NS, 2NS, 3NS conditions			
2NS	17.98***		
3NS	40.03***		
(ii) trials in 1NS, 2DS, 3DS conditions			
2DS	20.24***		
3DS	27.67***		
(iii) trials in 2DS, 2NS, 3DS, 3NS conditions			
interaction	-	-	-
length (3NPs)	16.01***	12.01***	44.31***
delay (NS)	-	-6.86*	-

Analyses of word durations allowed us to more precisely locate where in the phrase, the length and delay effects are observed. Among the numeral, color, and animal of NP1, only the animal word showed a significant effect of length. In NP2, both numeral and animal showed the effects of length, although the effects were larger in the animal word. A small yet significant length effect was also found in the conjunction *"and"*. In all these cases, the word durations increased with an increase in sentence

length. See Figure 3.20 and Figure 3.21, which shows the mean durations of each word (Figure 3.20) and plots the distributions of word durations that showed significant length and/or delay effects (Figure 3.21); Table 3.13 presents the statistical results. Given that the significant effect of length was observed at the edge of the NP – more specifically, the right edge of the NP (i.e. animal word), we can interpret that the participants were lengthening the end of the NP to plan for the upcoming part of the utterance.

Regarding the effect of delay, the durations of the conjunction *"and"* as well as the durations of the animal word of NP1 showed a significant effect, although the effect size was quite small in the durations of NP1 animal. See panels (a) and (b) of Figure 3.21. In both locations, the word durations were longer in the trials that had delayed stimuli. This suggests that the incorporation of delayed stimuli into an ongoing utterance indeed had started from the end of the first NP and continued throughout the interval between the first and the second NPs.

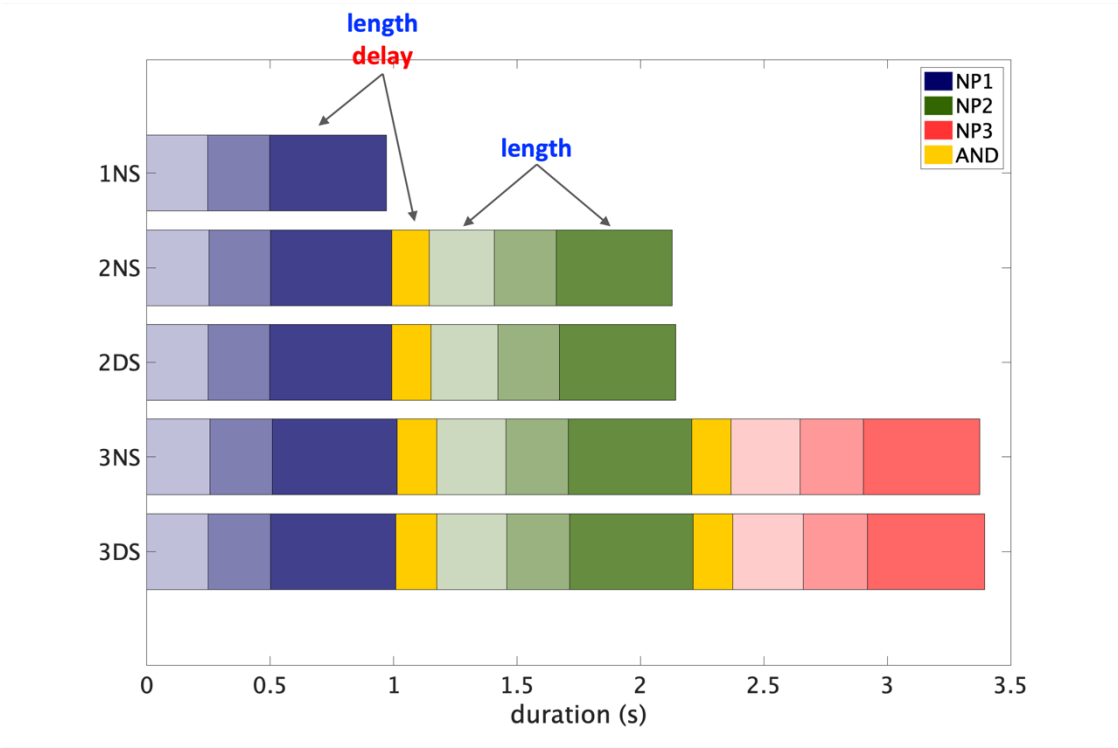

Figure 3.20. Mean durations of words within NPs and the conjunction "and", with markings of locations that had significant length/delay effects. Each of the blue, green, and pink bars indicates NP1, NP2, and NP3, respectively; within each NP, the lightest bar shows the average durations of numeral, the darkest shows the durations of animal, and the middle one shows the durations of color. The yellow bars show the average durations of "and". A significant effect of length was observed in ani1, AND1, num2, ani2 durs, and the effect of delay was found in ani1 and AND1 durs.

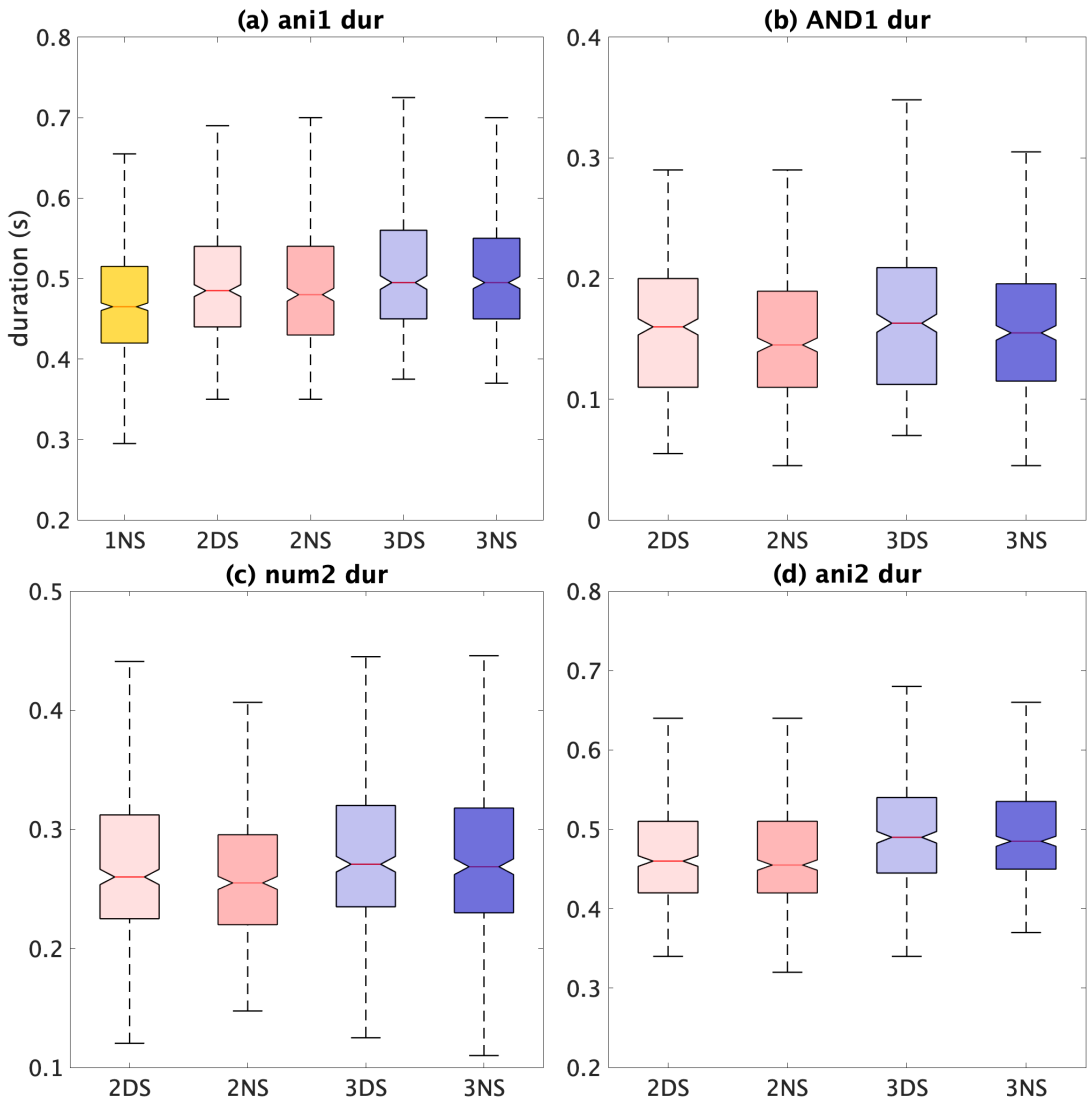

Figure 3.21. Distributions of ani1, AND1, num2, and ani2 durs, which showed the effect of length and/or delay, by experimental condition.

Table 3.13. Regression model coefficients of ani1 dur, AND1 dur, num2dur, and ani2 dur. The coefficients are in the unit of ms.

	ani1 dur	AND1 dur	num2 dur	ani2 dur
(i) trials in 1NS, 2NS, 3NS conditions				
2NS	15.41***			
3NS	32.39***			
(ii) trials in 1NS, 2DS, 3DS conditions				
2DS	22.86***			
3DS	31.82***			
(iii) trials in 2DS, 2NS, 3DS, 3NS conditions				
interaction	-	-	-	-
length (3NPs)	14.31***	8.53***	12.53***	29.45***
delay (NS)	-4.47*	-6.72***	-	-

3.4 Discussion

Overall, the results of the experiment first showed that participants set their F0 control parameters according to the initially planned sentence length. In particular, they set a higher F0 peak and valley as well as a wider F0 range when they had to produce longer sentences. It was also found that the participants were able to dynamically adapt to the changes in the sentence length that were cued after utterance initiation. When delayed stimuli appeared, participants lowered F0 peaks to a lesser extent than they did in the absence of the additional stimuli.

Several analyses were conducted to examine the F0 control hypotheses – i.e. *target* vs. *register*-control, and they provided support for *register*-control. Both variance and correlation analyses showed that the F0 peaks and valleys are not independent from each other but are correlated. The comparison of condition-prediction models, however, showed mixed evidence.

Sections 3.4.1 and 3.4.2 summarize and discuss the findings on the effects of initial

sentence length (3.4.1) and delayed stimuli presentation (3.4.2) on the F0 measures of the subject phrase. All of the analyses in this chapter were conducted on the participant-pooled data, yet differences may arise between participants in terms of how they control F0 variables according to the experiment manipulations. In this sense, a preliminary analysis was conducted to find out the inter-participant variations in F0 control, and the results are presented in Section 3.4.3. I also briefly discuss the properties of F0 contours of the four participants, whose data were excluded from the current analyses. Section 3.4.4 presents the results of the analyses which examined the main hypotheses of F0 control – *target* vs. *register*. Section 3.4.5 introduces some additional findings on F0 measures, specifically those at the end of the subject phrase and in the verb phrase. Lastly, Section 3.4.6 discusses results of the analyses on phrase and word durations, and what they inform us about the participants' speech planning mechanism.

3.4.1 Pre-planned F0 control

One of the important findings of the current experiment is that the F0 parameters of NP1 varied by the number of visual stimuli presented at the beginning of the trial. Specifically, F0 values of the peaks (P1) and the valleys preceding the peaks (Vpre1) as well as the ranges between the two variables (R1) increased in sentences with more initial stimuli. This suggests that participants made a pre-utterance plan considering the number of initial stimuli; thus, when they were aware that they had to produce a long utterance, they started from a higher F0 peak/valley or a wider F0 range. The motivation behind this control is to establish a sufficient tonal space for each utterance considering its length.

For the F0 measures of the initial NP, DS conditions patterned similarly to 1NS condition, which led us to code all these conditions (i.e. 2DS, 3DS, 1NS) with the same label (i.e. 1Pi) for the statistical analyses. This shows that in DS conditions, although more stimuli appeared after utterance was initiated, participants pre-planned for just a single NP, which demonstrates that our novel experimental paradigm worked as intended. It was a logical possibility that participants start all experimental conditions with the identical F0, after they learned that the delayed stimuli may show up in some cases. Namely, they may opt to start from a sufficiently high F0 peak/valley or a wide range in all conditions regardless of the number of initial stimuli (rather than adjusting F0 plan in the middle of the utterance in response to the delayed stimuli), yet this was not observed in our data.

One possible confounding factor in interpreting the initial F0 variations is that the silent rehearsal time (i.e. preparation time) varied according to the number of initial stimuli. Specifically, participants were given 2.7s for silent rehearsal when there was a single NP stimulus at the beginning, 4.4s for two initial stimuli, and 6.1s for three initial stimuli. As mentioned in Section 3.2.1, these periods were derived from the average sentence durations in the preliminary experiment data and were further tested on two native speakers of English who were naïve about the experiment. It is possible that a longer preparation time in two/three initial NPs has led participants to start from a higher F0 peak/valley or a wider range, as they have gained more respiratory energy during that time.

It is also important to point out that the initial F0 measurements did not vary significantly between trials in 2NS and 3NS conditions, although they differed in the

number of initial stimuli. Further analyses found that the two conditions differed only at the F0 peak (P1), and that the difference was also very small (around 1.31 Hz). This suggests that the participants simply distinguished one initial NP vs. more than one NP, not necessarily distinguishing all sentence lengths. This result is in line with the finding from Shih (2000), where participants produced three different pitch range variations when they were presented with ten variations of sentence length. Alternatively, it is possible that the three length variations of the current study (i.e. 1Pi vs. 2Pi vs. 3Pi) did elicit initial F0 differences, but the difference between 2Pi and 3Pi conditions was too small to detect with our statistical power. The results, however, still have observed the trend in the right direction, as the 3Pi trials had a slightly higher F0 peak than the 2Pi trials. A different possibility is that the length effect was determined by an expected number of phrasal units. Participants may have grouped two NPs into a single phrasal unit in both 2Pi and 3Pi conditions – i.e. 2Pi: [NP1 NP2], 3Pi: [NP1 NP2] [NP3], and that resulted little or no differences between the two conditions in the F0 measures of NP1.

As introduced in Section 2.2.3, the results of the previous studies were inconsistent on whether speakers raise their initial F0 peak according to sentence length. The current study adds to the literature that the length and the initial F0 are correlated, by showing that participants varied F0 parameters according to the number of initially presented stimuli. This study also informed us that it is not just the F0 peak, but other parameters such as F0 valley and range are controlled by speakers to reflect sentence length, which emphasizes the importance of analyzing F0 in more dynamic ways.

3.4.2 Adaptive F0 control

The other main finding of the current experiment is that the difference between the F0 measures of NP1 and NP2 varied by the occurrence of the delayed stimuli. Participants in general lowered P2 compared to P1, and the amount of reduction was smaller in the DS trials compared to the NS trials. The amount of reduction also varied by sentence length, as participants lowered F0 peaks to a lesser extent when they had to produce longer sentences. Combining these results together, we found that participants lowered F0 peaks to a lesser extent in sentences with three stimuli than those with two stimuli; and within each length, the amount of reduction was smaller in the DS trials than the NS trials. A significant effect of length was also observed in ΔF12, in which the participants compressed the F0 range of NP2 from NP1 in a lesser degree in longer sentences.

It is likely that the motivation for this manipulation is to reserve sufficient F0 space for the utterance, which is similar to the motivation for the speakers' initial F0 control that was discussed in the previous section. Thus, participants chose to lower the F0 peak or compress the F0 range from NP1 to NP2 only for a small amount in longer sentences, to make sure there is sufficient room of F0 for the remaining phrases. Likewise, as they encountered delayed stimuli unexpectedly, they adjusted their F0 control to make room for these new phrases, specifically by changing the amount of peak lowering from NP1 to NP2. The main drive for all these strategies is thus to have a sufficient F0 space until the end of the utterance and to avoid reaching the bottom of the register floor before the utterance ends.

The different adjustments of F0 peaks between DS and NS conditions have led the

F0 contours of the two conditions to become similar over the course of the utterance. The DS and NS conditions differed substantially at the initial NP, due to the number of stimuli presented at the beginning of the trial (i.e. 1Pi vs. 2Pi/3Pi (=2DS/3DS vs. 2NS/3NS), yet the difference between the two conditions was greatly reduced in the second and third NPs. The F0 values of the landmarks showed little difference between 2DS and 2NS conditions at NP2, and the difference between 3DS and 3NS conditions became smaller towards the end of NP3.

This finding in fact can be interpreted as participants having an abstract fixed pitch target for each NP location considering sentence length. For instance, participants have a hypothetical H pitch target at NP2 for sentences with two subject NPs, and likewise, they have an abstract H target at NP3 for sentences with three subject NPs. Our data then show that the participants were able to reach the hypothetical NP2 target in both 2DS and 2NS conditions (or reach the NP3 target in 3DS and 3NS conditions), although they started these conditions very differently. Given that the different adjustment of F0 across NP1 and NP2 between DS and NS conditions was found predominantly at the F0 peak measures, these results can be interpreted to support the *target*-control hypothesis. Yet, it should be noted that the similar interpretation can be made when we assume that participants have a fixed register ceiling for each of the NP location and sentence length.

In the current experiment, DS conditions always had a single NP at the beginning of the trial and one or two more NPs were presented after the initiation of the utterance. In sentences with three subject NPs, there was no DS condition, in which the two NPs were presented at the beginning and a single phrase was delayed. This was mainly due to the implementation challenge, as it was difficult to control the time to present the

single delayed stimulus. If the delayed stimulus is presented as soon as the utterance initiation is detected, as in other delayed stimuli, participants would have too much time to process the delayed phrase. Alternatively, one can present the stimulus after a fixed amount of time from the start of the utterance, but then at which point in the utterance the stimulus appears would differ a lot by participants depending on their speech rate. If, however, this design could be implemented, it would bring further insights about how speakers control their F0 according to the initial sentence length as well as the changes in the length. For instance, F0 landmarks of NP1 and NP2 could be compared between 2NS and the new 3DS conditions (i.e. NP1/NP2 (initial) + NP3 (delayed)), as both would have two NPs at the beginning of the trial. Moreover, comparisons could be made on the F0 landmarks of NP3 between the trials in 3NS vs. 3DS with a single initial NP vs. 3DS with two initial NPs. These comparisons would, however, be largely affected by when in time the delayed NP3 appears during production.

As far as I am aware, no previous studies have examined how speakers respond to changes in the utterance length that are made after utterance initiation. If we understand this study as an instance of the more general perturbation studies, the current work provides evidence that speakers are not only sensitive to the perturbations in the auditory feedback (Section 2.2.4) but also to the perturbations in the utterance plan itself – i.e. more fundamental changes in the content and length of the utterance. This finding is novel, but not surprising given that in spontaneous speech, speakers constantly update their utterance plan according to the internal and/or external factors. In this sense, one of the crucial contributions of the current study is that it developed a novel experimental paradigm which allowed testing of speakers' control system during production, without

using more online methods such as eye-tracking or EEG.

3.4.3 Inter-participant variations

In this section, I introduce a preliminary analysis on inter-participant differences in the variations of F0 with respect to the initial sentence length and the changes in the length.

Although seven participants, whose data were subject to detailed analyses, mostly increased initial F0 parameters in longer sentences and decreased F0 adjustments in the occurrence of delayed stimuli, the extent of the F0 variations across conditions differed by participant. As seen from Figure 3.6 which is introduced again in Figure 3.22, the difference between sentences with a single initial NP (1NS/2DS/3DS) vs. multiple NPs (2NS/3NS) was mostly observed in the F0 contours of the seven participants (first and second rows). They, however, differed in whether and how they distinguished 2NS and 3NS conditions. For instance, F0 values of P1 were higher in 3NS than 2NS in PA02, PA03, and PA04, but they were more similar in PA05 and PA07. On the other hand, the peaks of 2NS were indeed higher than 3NS in PA06, while the F0 differences between the conditions were overall small in the data of PA01. These observations are confirmed in Figure 3.23, which shows the distributions of P1 by condition in the data of PA02, 05, and 01. In this figure, the median of 3NS was higher than 2NS in PA02, yet they were more similar in PA05; the difference across conditions was smaller in PA01. It is also noticeable that the 3DS patterned a little differently from 2DS and 1NS (these conditions showed no significant differences in the main analysis) in PA05 and more significantly in PA01.

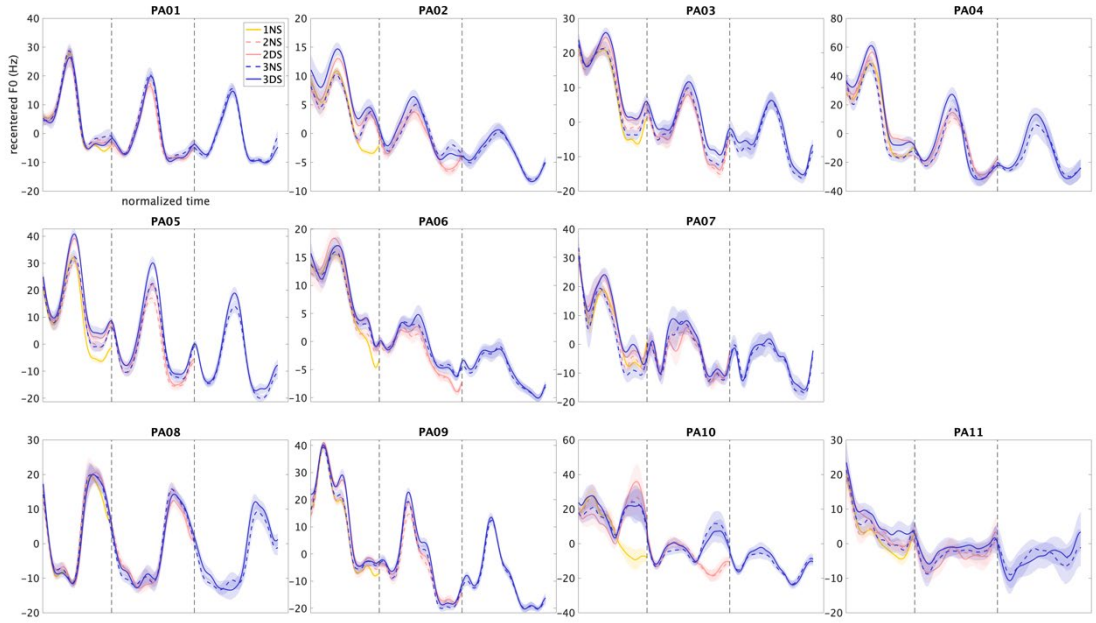

Figure 3.22. Smoothed/interpolated F0 contours of each experimental condition of each participant. See Figure 3.6 for the detailed information.

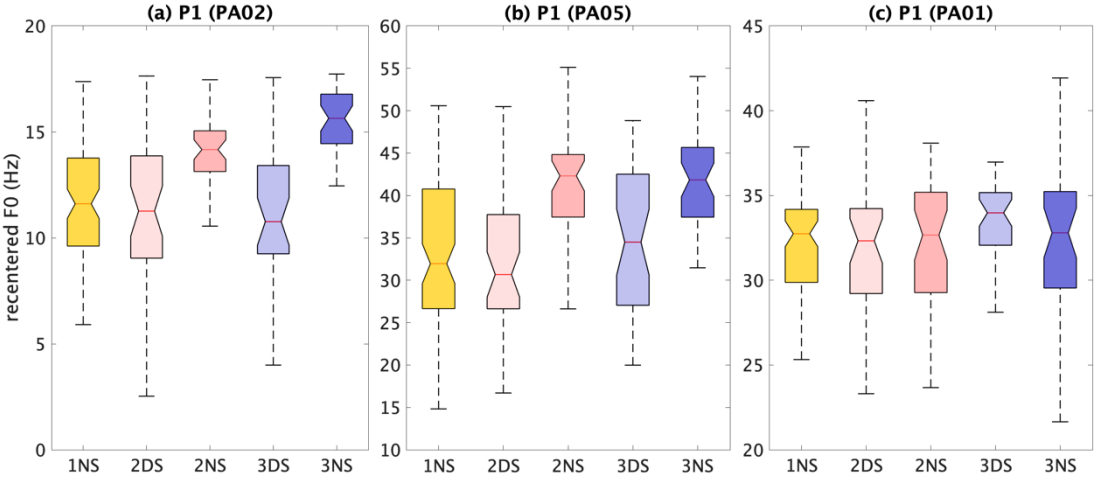

Figure 3.23. Distributions of P1 by experimental condition in (a) PA02, (b) PA05, and (c) PA01.

In addition, it was not just the extent of F0 differences, but the use of F0 parameters also differed by participant. Figure 3.24 shows the distributions of ΔP12 and ΔF12 in PA01 and PA07. In PA01 (top row), the difference between 2DS and 2NS conditions

was not very large in ΔP12, yet a larger difference between the two conditions was found in ΔF12 (a similar trend is also observed between 3DS and 3NS); this may suggest that PA01 more actively controlled F0 ranges to respond to the delayed stimuli. On the contrary, in PA07 (bottom row), the difference between 2DS and 2NS conditions was more substantial in ΔP12 compared to ΔF12, which suggests a more extensive usage of peak measures in PA07. Interestingly, in the ΔF12 of PA07 (Figure 3.24-(d)), there were a significant number of cases where the F0 range increased from NP1 to NP2 (i.e. $\Delta F12 < 0$) in the condition 2DS, compared to the condition 2NS. This shows that PA07 also chose to expand the range, instead of compressing it to a lesser extent, when they saw delayed stimuli.

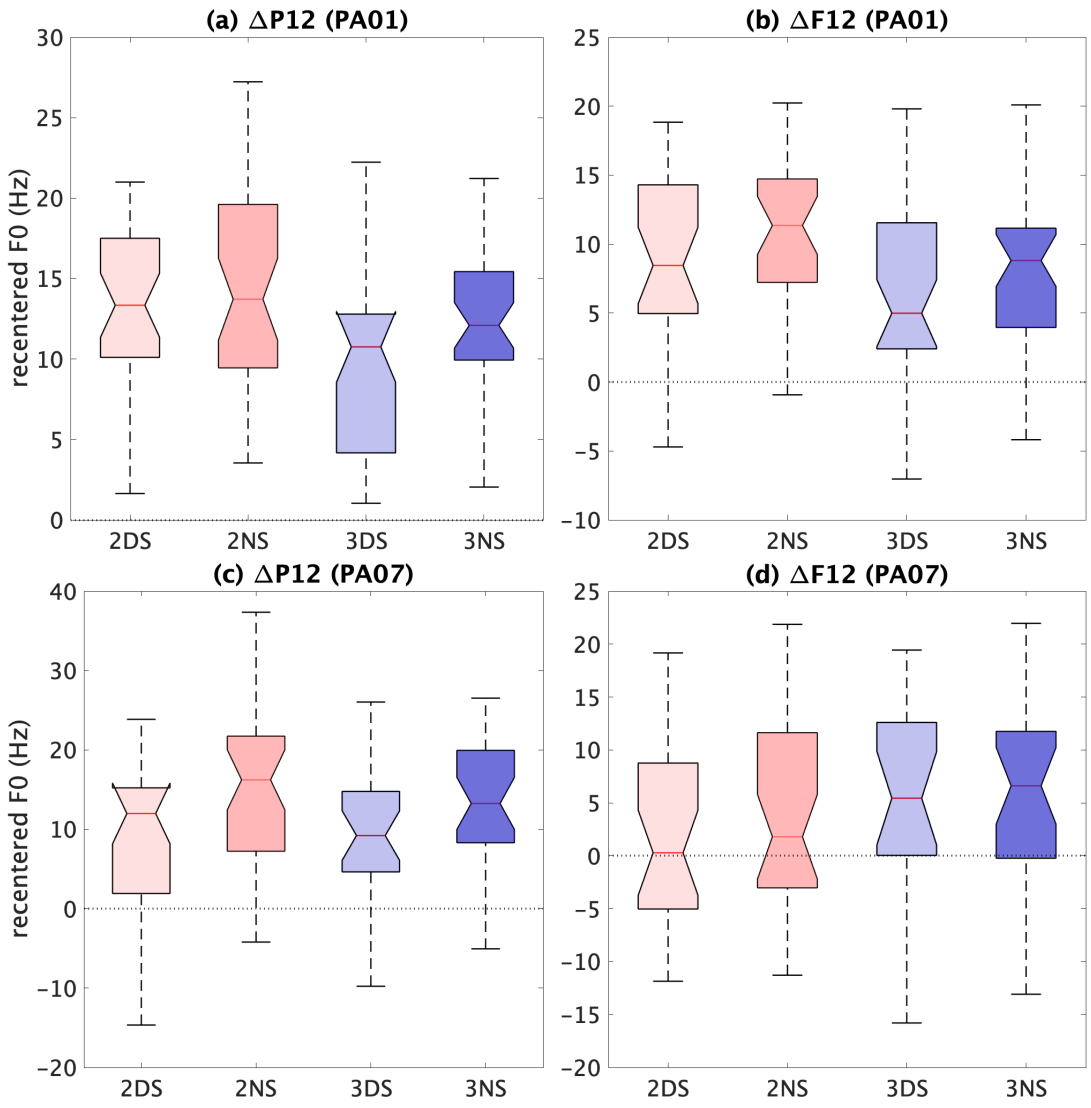

Figure 3.24. Distributions of ΔP12 and ΔF12 by experimental condition in (a)/(b) PA01, and (c)/(d) PA07. The horizontal dotted line marks 0, which shows that the F0 peaks or ranges were identical across phrases.

Some observations were also made on the data from four participants (PA08, 09, 10, 11 in Figure 3.22) which were excluded from the analyses. Specifically, F0 contours of PA08 and PA09 were worth investigating, as these participants showed consistent, dynamic F0 patterns throughout the experiment session, unlike PA10 and PA11. Figure 3.25 displays the average F0 contours of PA08 and PA09 from Figure 3.22 and the

distributions of the highest peak of NP1 by experimental condition; the F0 peak was specifically examined, as it was the F0 measure that showed the strongest difference between conditions in the main analysis. The boxplots in Figure 3.25 in fact did not show the pattern that was observed in the main analysis; namely, unlike the data from the seven participants, the difference between a single initial NP vs. multiple NPs was absent in these data. The median of 1NS was lower than the other conditions in PA08, but this resembles the markings of the end of the subject phrase, which was found in the Vpost measure in main analysis. For PA09, the difference between 2NS, 3NS and 1NS, 2DS, 3DS conditions indeed seems to emerge not at the most prominent (initial) peak, but rather at the second peak of NP1.

The lack of distinctions between conditions with a single initial NP vs. multiple NPs may be attributed to these data being noisier than the data in the main analysis. While segmental anchors were identified for each F0 landmark, and they were used to identify outlier measures in the main analysis, this procedure was not conducted when analyzing these additional data. In addition, it may be that the difference between conditions is present not at the highest F0 peak, but at other measures such as F0 valleys or ranges. Therefore, a more systematic investigation of these participants' data (and if possible, with more participants with similar patterns) is needed to find out how they control F0 parameters according to the experimental manipulations.

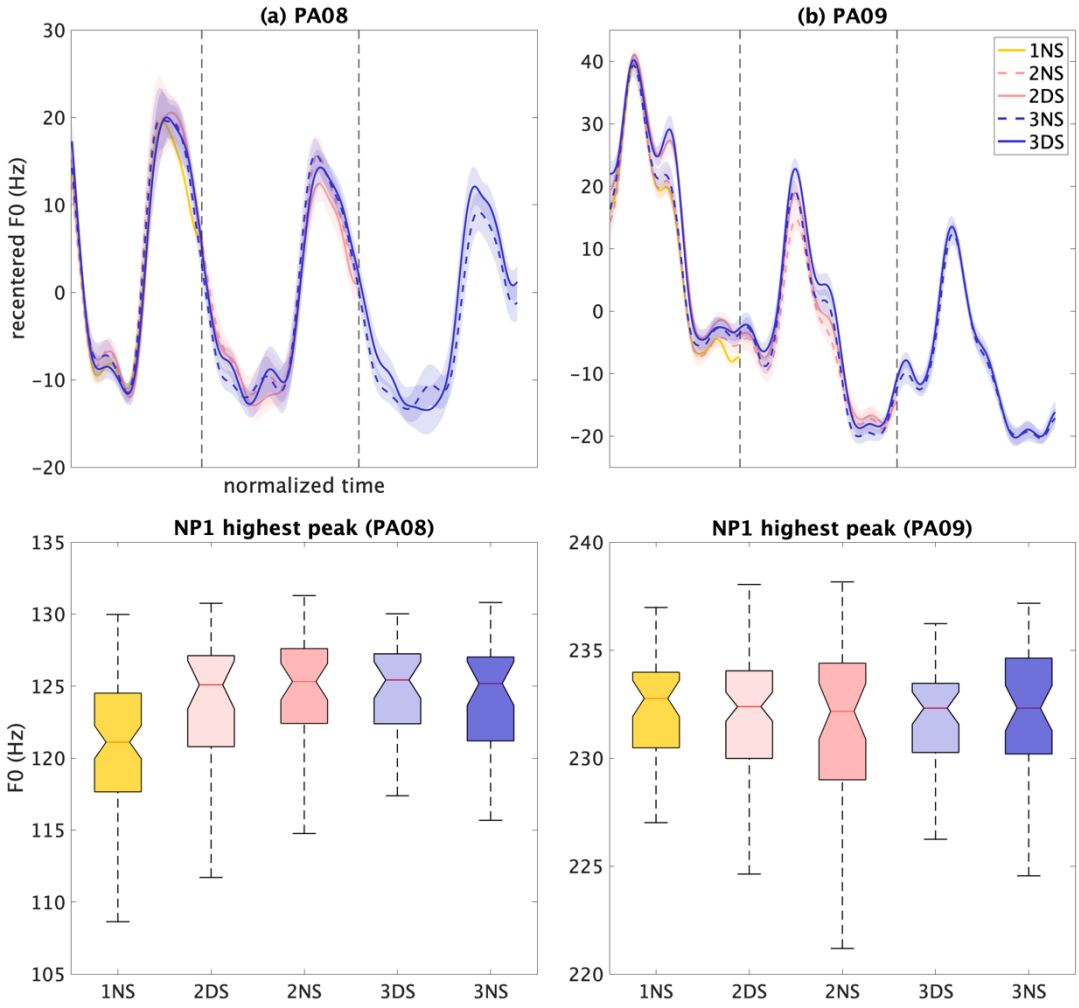

Figure 3.25. Smoothed/interpolated F0 contours of each experimental condition of (a) PA08 and (b) PA09. The bottom figures show the distributions of the highest F0 peak found in NP1 for the two participants.

The observations made in this section suggest that although we found an across-participant strategy of controlling F0 according to the initial utterance length and the changes in the length, the extent of using this strategy or F0 variable may differ across participants, which highlights the importance of inter-participant analyses. The analyses presented here are preliminary and qualitative, and the detailed analyses would provide further insights on how participants differ in their control of F0.

3.4.4 F0 control hypotheses

The F0 variations we have observed with respect to the initial sentence length and the changes in the length were subject to further analyses which aimed to examine the main hypotheses of F0 control – i.e. *target* vs. *register*-control hypothesis. Overall, the results provided supporting evidence for the *register*-control hypothesis. Three analyses were conducted on the F0 measures associated with NP1 and NP2: analyses of (i) the variance of F0 measures, (ii) the correlation between F0 peaks and valleys, and (iii) the comparison of NS/DS condition-prediction models. A summary of these analyses – the basic assumption, method, predictions from the *target* and *register*-control hypotheses, finding, supporting hypothesis – is presented in Table 3.14.

Table 3.14. Summary of the analyses conducted to examine the main hypotheses of F0 control (target vs. register). For each analysis, I introduce the assumption made for the analysis and the method. Then, the predictions from the target-control and register-control hypotheses, which were introduced in Table 3.1, are presented. They are followed by the finding and the hypothesis that is supported by the finding.

(i) variance of F0 measures	
assumption	if the sum of the variances of F0 peaks and valleys is larger than the variance of F0 ranges (calculated with peaks/valleys), it suggests that the peaks and valleys are not independently controlled.
method	for each NP/participant/condition, variances of peaks, valleys, and ranges were calculated and the sum of the first two was compared with the latter
target-control	$\sigma^2(\text{range}) \approx \sigma^2(\text{peak}) + \sigma^2(\text{valley})$
register-control	$\sigma^2(\text{range}) \ll \sigma^2(\text{peak}) + \sigma^2(\text{valley})$
finding	$\sigma^2(\text{range}) \ll \sigma^2(\text{peak}) + \sigma^2(\text{valley})$
supporting hypothesis	register-control

(ii) correlation between F0 measures	
assumption	when peaks and valleys within NP are highly correlated, register span is the control parameter
method	for each NP/participant/condition, correlation between F0 peaks and valleys was calculated
target-control	N/A
register-control	peaks and valleys within NP are highly correlated
finding	peaks and valleys within NP are highly correlated
supporting hypothesis	register-control

(iii) model comparison	
assumption	the predictor of the model with the lowest AIC is the control parameter
method	for each NP, compared three logistic regression models on how well they predict DS/NS conditions with AIC values – the model had either F0 peaks, valleys, or ranges as a predictor
target-control	range model > peak, valley models
register-control	range model < peak, valley models
finding	NP1: peak < range < valley NP2: range < valley < peak
supporting hypothesis	target-control (NP1), register-control (NP2)

First, in the variance analysis, at each NP, the sum of the variances of F0 peaks and valleys was compared with the variance of ranges for each participant and condition.

As presented in Table 3.14, if there is a large amount of inequality in the sum of the inflection point (peaks and valleys) variances and the variance of their differences (ranges), it shows that the inflection points (estimates of targets) are not independently controlled. The results indeed found that, in most participants and conditions, the variance of the range was substantially smaller than the sum of the variances of the peaks and valleys in both rises/falls and NP1/NP2. This inequality to some extent could be viewed as the evidence for *register*-control. There of course could be some other unknown mechanisms that cause the H and L targets to covary, but it is difficult to imagine what that mechanism would be other than the register span.

An additional observation was that the difference between the variance of the range and the sum of the variances of peaks and valleys was relatively smaller in the comparison between P+Vpre and R at NP1. This may seem to suggest that the peaks and the preceding valleys at NP1 are less correlated than other measures. Yet, I believe that this result rather suggests that the Vpre1 was not a good reflection of the bottom of the register; see Figure 3.11 in Section 3.3.1.1, where the Vpre1 was not close to the register floor. Thus, since the range from Vpre1 to P1 did not reflect the span as well as the fall (the range from P1 to Vpost1, the latter of which was closer to the register floor), a relatively weak correlation between the peaks and the preceding valleys was observed at NP1.

Second, analysis was conducted to find out whether peaks and valleys following the peaks, the ranges (F0 falls) of which were considered to estimate the register span, are correlated within each NP. A positive correlation was found between the peaks and the valleys. When the R-squared values were examined, a moderate to high correlation

was found between the two measures in many participants and conditions, providing evidence for the *register*-control hypothesis. The correlation between the two measures, however, could be equally predicted from the *target*-control hypothesis, if we assume that speakers control both H and L pitch targets in a correlated way. Yet, the *register*-control is a better account, since an additional assumption that speakers control H and L targets in a correlated manner, is required for the *target*-control interpretation. Thus, although both *target* and *register*-control hypotheses predict a high correlation between peaks and valleys, the result is interpreted here as the evidence for *register*-control (cf. the prediction of the *target*-control hypothesis is marked as N/A in Table 3.14).

Besides within-NP F0 measure correlations, the correlations between F0 peaks, valleys, and ranges across NP1 and NP2 were also examined (i.e. P1-P2, Vpre1-Vpre2, F1-F2); however, their results were not introduced in this chapter, as no systematic pattern was found in the data. For this analysis, the correlation of DS trials was compared with that of NS trials within each participant, based on the assumption that the correlation of F0 measures across NPs would be more robust in the NS trials compared to the DS trials. This is because the speakers' F0 control is interrupted by the presentation of the delayed stimuli in the DS trials, while speakers can proceed with their pre-utterance F0 plan in the NS trials. Yet, the correlation values were overall very small, and no systematic pattern was observed in the data.

Lastly, the model comparison results suggested different F0 control parameters for each NP. In NP1, the AIC value was lowest in the model that had F0 peaks as the predictor, which was followed by the model that had ranges as the predictor, and then the valleys (AIC: H < SP < L); on the other hand, in NP2, the AIC value was lowest in

the model with F0 ranges as the predictor, then valleys, and then peaks (AIC: SP < L < H).

These results can be interpreted as speakers controlling H targets at NP1 and register span at NP2. The control of H targets at NP1 may reflect a unique property of utterance-initial F0 control, or it may be in fact caused by the variation in the initial sentence length (i.e. the specific manipulation of the current experiment led participants to control H targets at NP1 rather than other parameters). Yet, there are a couple of weaknesses in this interpretation. The first is that we do not have any theoretical basis to argue that speakers control different F0 parameters at different phrases. The second is that the time that the delayed stimuli are incorporated into the ongoing utterance can be a confounding factor in interpreting the model comparison results. At NP1, we found that the early F0 measures (i.e. Vpre and P) were not affected by the occurrence of delayed stimuli (i.e. they were affected only by the initial sentence length). However, it is not clear whether this is also the case for Vpost (which affects the register span) that occurred later in the phrase, as it is possible that participants may have already started incorporating the delayed stimuli into their utterance around this point. That said, comparing three models on how well they predict the delayed vs. no-delayed stimuli conditions especially in NP1 may not be ideal, considering the different effects of delayed stimuli on the F0 measures.

In sum, the F0 variations we have observed with respect to the experiment manipulations showed evidence for the *register*-control hypothesis, with some partial evidence for the *target*-control in the model comparison results. It is, however, crucial to emphasize that since F0 peaks, valleys, and ranges are merely the *estimates* of pitch

targets and pitch register, it is difficult to find irrefutable evidence for the F0 control hypothesis. The reason is that our interpretations are based on the assumption that the pitch targets are never undershoot/overshoot and are likely to be at the edges of the pitch register, and thus, the surface F0 measures such as peaks/valleys and ranges properly represent the targets and register.

Note that the *register*-control hypothesis is also supported in the modeling study in Chapter 4 yet with different types of evidence; in the modeling, pitch targets and register are no longer estimated through surface F0 measures (peaks/valleys/ranges) but are inferred from optimization. The details of the computational modeling will be provided in the next chapter.

3.4.5 Additional findings on F0 control

Interesting patterns were observed from the investigations on other F0 measures as well. The first is that the participants marked the end of the subject phrase with F0, specifically through F0 valleys following the peaks (Vpost), and in some cases, F0 falls (F). In particular, when the given NP was the final NP of the subject phrase, F0 values of Vpost were particularly low, which were often accompanied by a large F. Thus, at NP1, 1NS trials had the lowest Vpost, while at NP2, the Vpost values of 2NS and 2DS trials were lower than those of 3NS and 3DS.

An extra lowering of Vpost in the subject final NP may be the phonetic effects, in which it is the combination of the declination over the subject phrase and the declination over the entire utterance, as in the finding of "declination within declination" in Ladd (1988).

An alternative explanation can be made via prosodic phrasing. Phonologically, each subject NP would be analyzed as having an L+H* accent, which is manifested as Vpre and P; the phonological origins of Vpost, however, are less straightforward. One possibility is that it is the phonetic realization of a low phrase accent (L-); this may arise in a case when each s subject NP constitutes an intermediate phrase – for example, $[[NP1]_{ip}\ [NP2]_{ip}]_{IP}\ [[VP]_{ip}]_{IP}$ (ip: intermediate phrase, IP: intonational phrase). Another possibility is that it is the realization of a low boundary tone (L%), in which case, the prosodic phrasing would be $[NP1]_{IP}\ [NP2]_{IP}\ [VP]_{IP}$. It is less likely, but Vpost can also be the realization of a low pitch accent (L*), which is anchored to the second syllable of the animal word; it may arise in 1NS condition where the subject phrase and VP are grouped together as $[[NP\ VP]_{ip}]_{IP}$ or in 2NS conditions where the subject NPs are grouped together as $[NP1\ NP2]_{ip}$, especially in fast speech rate. It is difficult to determine the phonological identity of this landmark before analyzing productions of individual trials, as prosodic phrasing may well differ within and across participants and experimental conditions. Yet, considering the syntactic structure of the current stimuli, it is likely that the Vpost is the realization of a low phrase accent or boundary tone.

Our finding that the NP ends with a particularly low F0 can thus be interpreted either as a prosodic category shift or gradient changes within the category. First, when the given NP is the final NP of the subject phrase, participants may produce that NP as a larger phrasal unit. For instance, when the given NP is followed by another NPs, it is produced as an ip, but when it is subject-final, it is produced as an IP. Second, the prosodic phrasing of the subject NPs is same regardless of the presence of the following NPs – e.g. the NPs with or without the following NPs are produced identically as an ip

or IP, yet they exhibit gradient phonetic differences. Further work on the prosodic structure of the data would provide more information on the phonological nature of the Vpost measure and the subject-final markings. Moreover, prosodic phrasing analyses would also demonstrate how participants organize subject NPs in their pre-utterance plan as well as how their phrasing is affected (interrupted) in the occurrence of delayed stimuli.

The second interesting finding comes from the F0 measures of VP. The F0 values of VPmax and VPmin were examined, and the analyses found a significant effect of length in both F0 measures. Specifically, F0 values of VPmax and VPmin were higher in shorter sentences. This suggests that the participants started and ended the verb phrase with a higher F0 in sentences with fewer subject NPs.

The effect of length on VPmax is not surprising, given the location of VP within an utterance. In 1NS trials, VP is the second phrase from the utterance start, while in 2NS/2DS trials, it is the third phrase, and in 3NS/3DS trials, it is the fourth phrase. Considering F0 declination or downstep, it is expected that the VPmax is higher in 1NS trials, as there are fewer intervening phrases from the start of the utterance to the VP.

An interesting result comes from VPmin, which also exhibited a length effect. Our result contrasts with a well-known finding that the utterance-final F0 is relatively stable for a given speaker. For instance, Liberman and Pierrehumbert (1984) found that the utterance-final F0 values do not vary with sentence length or the type of the stimuli, which led them to argue that the utterance-final low F0 is an invariant characteristic of a speaker's voice; similar results were found in Maeda (1976), Boyce and Menn (1979) for English, and Prieto (1996) for Spanish. Yet, there were also studies which argued

against the near-constancy of the utterance-final F0; for instance, Ladd and Terken (1995), Shriberg et al. (1996), Thorsen (1980), and Hirschberg and Pierrehumbert (1986). The current result is in line with the latter studies, providing evidence that the utterance-final F0 can differ by factors such as sentence length.

3.4.6 Speech planning evidenced by durations

Analyses of the phrase durations found a significant effect of sentence length in NP1, NP2, and NP1-NP2 interval durations. In all cases, durations increased when there were more NPs in the subject phrase, which is similar to the finding from Sternberg et al. (1988). Further analyses of the word durations found a significant effect of length in ani1, AND1, num2, and ani2 durations. The effect of stimulus delay was found in NP1-NP2 dur at the phrase-level and ani1 and AND1 at the word-level. I will first discuss the length effect that was observed in NP1 and NP2 (ani1, num2, ani2) and then discuss the effects of delay (and length) observed in the interval between NP1 and NP2 (ani1, AND1).

3.4.6.1 NP1 and NP2 durations

Our data showed that the durations of NP1 and NP2 were longer in sentences with more subject NPs. The participants in our study were given a sufficient time to prepare their production and were explicitly instructed to silently rehearse the sentence before they initiate an utterance. It is possible that this explicit preparation stage could prevent the length effect from occurring, if we assume that the participants memorize the whole utterance and produce it as soon as they see the start signal. Yet, this was not the case;

even with the preparation, the durations of the phrase differed by sentence length. This shows that the participants' production of the given NP was affected by the presence of the following NPs; more specifically, while they were producing the first NP, the presence of the second NP (and maybe also third NP) influenced their production of the first NP, and likewise, when they were producing the second NP, the presence of the third NP affected their production of the second NP. I argue that the participants were planning for the upcoming NPs while producing the current one, and that caused the significant effects of length in NP1 and NP2.

Another logical possibility is that the prosodic phrasing differed by sentence length. For instance, in 2NS trials, NP1 is produced as an intermediate phrase, while in 3NS trials, the same NP is produced as an intonational phrase, and thus, the amount of phrase-final lengthening is reflected as the length effect. Note, however, that this explanation is not mutually exclusive from the one given above, as participants could produce the NP of the same location into different prosodic categories to gain more time to process the upcoming part of the utterance.

An interesting finding regarding the length effect is that the effect was much larger in the durations of NP2 than in NP1. The regression coefficient of the condition with the three NP stimuli compared to the two NP stimuli was 16.01 ms at NP1, while it was 44.31 ms at NP2. This suggests that participants needed more time to plan for NP3 at NP2 compared to the time needed to plan for NP2 at NP1. One reason for this difference is that the participants did not plan for NP3 before utterance initiation. Alternatively, it is possible that they did make a plan, but they needed more time to retrieve and process NP3, as some time had passed after setting up the initial plan.

Further analyses of word durations found that the length effect observed in NP1 and NP2 was indeed derived from the lengthening of the animal word. This suggests that the participants were lengthening particularly the right edge of the phrase while they were planning for the upcoming part of the utterance. Note that a small yet significant effect of length was observed in the numeral of NP2, which proposes a possibility that participants lengthen both edges of the phrase to plan for the following NPs. The length effect, however, was not observed in the numeral of NP1, presumably because it was the very beginning of the utterance, and the silent rehearsal time was explicitly given right before the utterance start. Further experiment with the sentences that have more subject NPs would show whether it is only the right edge of the phrase that is used for planning, or the planning occurs at both the left and right edges.

3.4.6.2 NP1-NP2 interval durations

A significant effect of delayed stimuli presentation was found at the interval between NP1 and NP2. In particular, the interval durations were longer in sentences with delayed stimuli (DS > NS conditions). In the analyses of word durations, not only the conjunction *"and"* showed a significant effect of stimulus delay, but also the animal of NP1 showed the effect (although the effect was small). The results show that the lengthening started from the final word of the first NP and continued (indeed more heavily lengthened) throughout the interval between the first and second NP. In other words, participants started incorporating the delayed phrases into their utterance, from the end of the phrase that was initially presented, but more extensively after that phrase was uttered. The length effect was additionally observed at NP1-NP2 dur and AND1 dur, which suggests that it took more time for participants to incorporate two delayed

NPs compared to a single NP.

Similar evidence was also found in the analyses of the F0 measures. A significant effect of delay was found in all NP1 measures as well as the F0 valleys preceding the peaks at NP2. After Vpre2, the F0 contours of 2NS and 2DS trials became almost identical, and the difference between the contours of 3NS and 3DS trials was also small, which suggests that the presence of delayed stimuli no longer affected these landmarks; this is in line with the lack of delay effect in phrase/word durations of NP2. We can thus argue that the processing for the delayed stimuli starts from the end of NP1 and lasts until the beginning of NP2, although the main region that they are processed is the interval between NP1 and NP2.

It is, however, important to point out that the incorporation of delayed phrases at the end of or after NP1 may have arisen due to the way the length was manipulated in the experiment. Since the current study varied sentence length in the unit of phrase, it may have led participants to plan or process this unit as a whole, thus preventing the effects to arise within the phrase. When especially considering that the delayed stimuli were presented on average 86.4 ms after utterance initiation (Section 3.2.2), which was indeed almost immediately after the start of the production, the delay effects emerged fairly late as they were found around the end of the initial NP (cf. the duration of NP1 was on average around 1s as shown in Figure 3.18).

This may suggest that participants needed some time to process the delayed stimuli and reflect them in their production, yet it is also possible that the effects appeared at this location as this was the end of the phrasal unit. In other words, it is because participants processed the sentence phrase-by-phrase, the delay effects showed up once

a given phrase was uttered; that said, if length was varied in other syntactic/prosodic units (e.g. syllables, words), the effects of delayed stimuli presentation may have been found earlier in the utterance. Further studies that involve different length manipulations – for instance, varying different syntactic units (e.g. syllables, words, sentences) or structures (e.g. branching vs. non-branching) – would bring interesting insights on how changes in the length of the utterance are incorporated online.

Overall, assuming that the longer phrase/word durations are the reflections of the longer processing time, our data showed that the participants constantly plan for the upcoming part of the utterance during production. Moreover, they were able to adapt to the changes in the sentence that were made after utterance initiation by lengthening the interval between the initial NP and the newly presented NPs; note that in F0 analyses, the adaptation was reflected in the participants' adjustment of F0 between the first and the second NPs. Together with the finding that participants made a pre-utterance F0 plan, the experiment data of this chapter overall showed that participants are sensitive to the properties of the sentence that they are going to produce *before* utterance initiation as well as *during* production.

CHAPTER 4
COMPUTATIONAL MODELING

4.1 Introduction

In the previous chapter, we have observed that participants vary F0 parameters (i.e. F0 peaks, valleys, ranges) according to the initial sentence length as well as the changes in the length that were made after utterance initiation. This provided evidence that speakers make a pre-utterance plan that takes sentence length into account, and if necessary, they can dynamically adjust the plan online. In addition, analyses of F0 control mechanism found that these F0 variations are likely to arise from the control of pitch register. The modeling study in the current chapter aims to further examine the F0 control hypotheses (i.e. whether it confirms the *register*-control or rather supports the *target*-control) with a different type of evidence.

For this purpose, I introduce the gestural model of F0 control that is developed in the framework of Articulatory Phonology (AP). The biggest advantage of the modeling is that pitch targets and register are estimated through optimization, unlike in the analyses in the previous chapter, in which the targets and register were inferred from the surface measures of peaks/valleys and ranges. Given an empirical F0 contour, the parameters of the model (including the ones that are associated with targets and register) are optimized to minimize the root mean squared error (RMSE) between the model-generated contour and the empirical one. It is true that even with optimization, the estimates of pitch targets and register are not direct (in fact, we can never observe these control parameters directly), and they may or may not represent actual targets/register

that are employed by speakers; yet, the modeling is expected to provide a novel form of evidence to examine the main hypotheses of F0 control.

The modeling was conducted on a smoothed and interpolated F0 contour of the subject phrase. Given that each subject NP had one F0 peak, which was preceded and followed by an F0 valley (i.e. F0 valley – F0 peak – F0 valley), I posited one high (H) and one low (L) F0 gesture for each NP. The main feature of the current F0 model is that the target parameters of these H/L F0 gestures are always specified in a normalized coordinate (in the interval from 0 to 1), and they are mapped to actual F0 values through another set of parameters that correspond to pitch register. Thus, pitch targets are no longer described in the unit of Hz (as in Chapter 3) but are described in a more abstract sense. In addition, while we had to assume that the H and L targets (F0 peaks and valleys) are at the edges of register (or at least highly correlated) to infer register span from the measures of F0 ranges, that assumption is no longer necessary; rather, the optimization algorithm reconstructs the tonal space for a given F0 contour, in which the targets may or may not be located at its edges.

Different versions of F0 models were constructed, including those that reflect our main hypotheses of F0 control – i.e. *target* vs. *register*-control. Specifically, F0 models differed by whether gestural target values and register parameters were defined phrase-specifically (i.e. by each subject NP) or utterance-specifically[5] (i.e. shared across

[5] In this chapter, "utterance-specific" refers to the setting of the model where gestural targets or register parameters are defined at the level of subject phrase. This means that all NPs within the subject phrase share the same gestural target values or register parameter values. Since the subject NP is not the whole utterance (i.e. VP is excluded), it may be misleading to refer to this as "utterance-specific" (but rather refer it as "subject phrase-specific"), yet this expression was adopted to distinguish this setting from the setting where parameters vary by the conjoined subject NP.

subject NPs). In the model that assumed *target*-control, gestural targets were allowed to vary from phrase to phrase, while pitch register remained constant throughout the utterance; on the other hand, in the model that assumed *register*-control, register parameters were allowed to vary by phrase, whereas the values of the gestural targets were invariant across phrases. Note that the latter model, which assumed invariant gestural targets, is more consistent with the early description of gestures in AP (e.g. Browman & Goldstein, 1990a, 1990b). See Figure 4.1 for schematic illustrations of the *target*-varying (left figure) and *register*-varying (right figure) models. Besides these models, F0 models in which both gestural targets and register parameters were defined phrase-specifically or utterance-specifically were also tested. See Table 4.1 for the full list of F0 models.

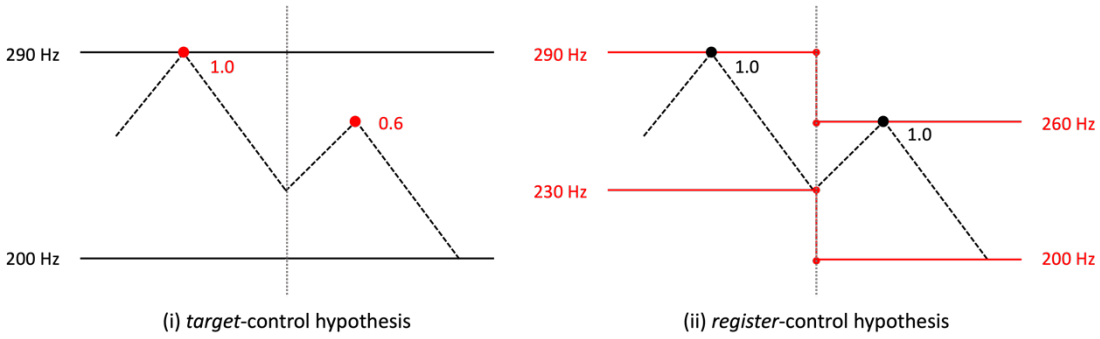

Figure 4.1. Schematic representations of F0 models that assume target (left) and register (right) control hypotheses. See Figure 1.2 for the detailed information about the figures. Note that these are the schematic illustrations; in the experiment, the smoothed and interpolated F0 contours were modelled. In the F0 model that exemplifies the target-control hypothesis, gestural target values (but not register parameters) vary by phrase, while in the F0 model that exemplifies the register-control hypothesis, the register parameters (not gestural targets) are varied.

Two types of modeling experiments were conducted. In the first experiment, which I refer to as the "speaker-level modeling", F0 models were tested on the average

time-warped F0 contours that were made for each speaker and experimental condition. In the second experiment, which is referred to as the "trial-level modeling", the models were tested on the F0 contours of all individual trials. To identify which F0 model shows better performance, the RMSE between the input F0 contour and the contour generated by the model (i.e. model cost) was examined. The model with a lower cost (i.e. smaller difference between the optimized contour and the input contour) was considered to provide better fits.

In sum, four F0 models were tested in the study, which varied by whether gestural targets and register parameters are defined utterance-specifically or phrase-specifically (i.e. by-utterance vs. by-phrase). Table 4.1 presents the list of F0 models tested in the experiment along with specific predictions of our main hypotheses of F0 control. Among these models, the main focus of the comparison was Model 2 and Model 3, each of which reflects the *target*-control and *register*-control hypotheses. The *target*-control hypothesis would predict that the cost of Model 2 would be lower than the cost of Model 3, whereas the *register*-control hypothesis would predict the contrary result (cost: Model 2 > Model 3). It is also expected that Model 1 would show poor performance, while Model 4 would perform well, given that the model with more phrase-specific parameters (thus with more free parameters) would better reflect the specific properties of F0 contours of each phrase. Yet, it is likely that Model 4 is overly powerful, and the same quality of model fit can be derived from the model which allows less variation (i.e. Models 2 or 3). These predictions are identical in both speaker-level and trial-level modeling experiments.

Table 4.1. F0 models tested in the current study and the predictions of the F0 control hypotheses. The top table lists the four models: "by-utterance" means that the gestural targets or register parameters are defined utterance-specifically, and "by-phrase" means that the parameters vary by phrase. The bottom table introduces the predictions of the experiments under each hypothesis of F0 control.

F0 models	**target**	**register**
Model 1	by-utterance	by-utterance
Model 2	by-phrase	by-utterance
Model 3	by-utterance	by-phrase
Model 4	by-phrase	by-phrase

Predictions	
target-control hypothesis	cost: Model 2 < Model 3
register-control hypothesis	cost: Model 2 > Model 3

To preview the results, the "*register*-control model" overall fit the empirical data better than the "*target*-control model". In both experiments, the cost of Model 3 was lower than the cost of Model 2, suggesting that the speakers control pitch register to produce variations in F0. This means that speakers have one set of abstract cognitive representations of high and low throughout an utterance, but they control tonal space to realize these abstract representations into different F0 peaks and valleys.

In the rest of this section, I present a brief introduction of AP (specifically with respect to F0), which is the theoretical framework that the gestural model is based on. Section 4.2 details the basic mechanisms and implementation details of the gestural model and introduces various F0 models and two modeling experiments. Section 4.3 provides further details on the experiment methods. Section 4.4 presents the experiment results, which are further discussed in Section 4.5.

4.1.1 Articulatory Phonology and F0

Within the framework of Articulatory Phonology (AP), "gestures" are the fundamental units of speech, which are simultaneously the units of cognitive representation and the units of physical action. Gestures are characterized as dynamical systems (Saltzman & Munhall, 1989), which are defined by parameters such as gestural target and stiffness (related to the time it takes to reach a target). The presence/absence of gestures, their respective locations, and the values of the associated parameters provide a basis of phonological contrasts as well as physical realizations.

The gestures specify "vocal tract variables" – i.e. where in the vocal tract the constriction must be present, how much degree of constriction it should be, and for some gestures, how large the aperture should be. A total of eight tract variables were proposed in Browman and Goldstein (1989): they are lip aperture (LA) and lip protrusion (LP), tongue tip (TT) and tongue body (TB) constriction location (CL) and degree (CD) (i.e. TTCL, TTCD, TBCL, TBCD), and velic (VEL) and glottal aperture (GLO). These tract variables are associated with model articulators such as lips and jaw. Each tract variable is modelled with a critically damped, second-order differential equation; when gestures that are associated with a tract variable become active, they change the equilibrium or target of the tract variable.

The notion of the gestures and tract variables has been extended to suprasegmental features as well. Specifically, Gao (2008) proposed pitch gestures[6] – high (H) and low

[6] I avoid the use of the term "tone gestures" and prefer to use "F0 gestures" or "pitch gestures" in this chapter. This is because these gestures not only model lexical tones, but also intonational tones such as pitch accents and boundary tones.

(L) gestures – to model lexical tones of Mandarin Chinese. Unlike constriction gestures, physiological mechanisms involved in F0 production cannot be easily described in geometric coordinates, as the lungs, trachea, and larynx as well as various muscles coordinate to make vocal folds vibrate. In addition, while we can precisely define the movement of articulators for oral constrictions in an absolute sense, pitch is understood to be a more relative concept – i.e. for instance, while we can specify where to place our tongue tip or how large the lip opening should be, it is not the case that speakers are locating H or L targets at certain specific frequencies. Modeling of the pitch gestures, therefore, was done more at an abstract level. Gao (2008) stated that "the fundamental frequency is treated as the goal of the tone gesture. Our aim is simply to model the dynamics of the goal variable (f0) itself, rather than the control of the physiological articulators (such as CT (cricothyroid) and sterno-hyoid distance) responsible for f0 variations. This means that f0 is effectively an articulator in the model as well as a goal variable" (p.42).

Yet, F0 gestures/tract variables are not so much different from oral constriction gestures/tract variables, if we consider that the coordinate dimension that the F0 is defined is simply the F0 coordinates, not the coordinates that index the vocal tract geometry as in lips or tongue variables; pitch gestures, in this sense, are also modelled as a second-order differential equation. Note that the H and L pitch gestures differ from the H and L tonal targets in the Autosegmental-Metrical (AM) intonational phonology: tonal targets in AM refer to the turning points (events) in the surface F0 contours, such as F0 peaks and valleys, while F0 gestures in AP are dynamical systems that are turned on, reach targets, and are turned off, and are defined with parameters such as gestural

onset, duration, and target. See Figure 4.2 for the comparisons of H/L tonal targets in AM and H/L pitch gestures in AP.

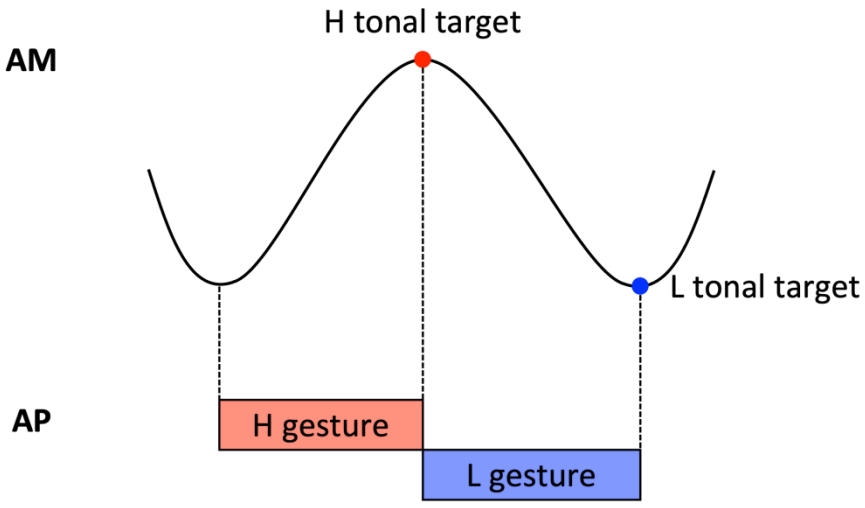

Figure 4.2. Comparisons of H/L tonal targets in the Autosegmental-Metrical (AM) intonational phonology and H/L F0 gestures in the Articulatory Phonology (AP). The black solid line shows a sample F0 contour. F0 peaks and valleys (red, blue circles) are the H and L tonal targets in AM, but they indicate the timepoints that the F0 gestures (red and blue boxes) are turned on and off. Although not shown in this figure, F0 gestures may overlap with each other.

With the introduction of pitch gestures, studies have examined the coordination between F0 gestures and constriction gestures. For example, Gao (2008) found that the pitch gesture that represents lexical tone in Mandarin Chinese behaves like an additional consonantal gesture, which may affect the within-syllable timing of the consonant (C) and vowel (V) constriction gestures. On the other hand, F0 gestures that are associated with pitch accents did not alter the timing of C and V within a syllable; these results were observed in Mücke et al. (2012), who examined the alignment pattern of bitonal pitch accent in Catalan and Viennese German, and Niemann et al. (2011), who examined the pitch accent timing in Italian and Standard German. The F0 gestures have also been

extended to model boundary tones; however, the coordination between boundary tone and C and V gestures differed by studies. For example, Katsika et al. (2014) found that the boundary tone gesture does not modify the CV coordination as pitch accents, in line with their post-lexical status, while Yi (2017) found the opposite results.

Besides F0 gestures, many efforts have been made over the last decades to incorporate prosody into the framework of AP. One example is the π-gesture model proposed by Byrd and Saltzman (2003). The π-gesture is activated at prosodic boundaries and slows down the unfolding of the gestures that are coactive. The π-gesture readily accounts for boundary-related effects such as lengthening of boundary-adjacent gestures, less temporal overlap between the gestures, and an increase in the magnitude of articulatory movement, which have been extensively demonstrated in many studies on various languages. Another example is the μ-gesture which was proposed to account for the prosodic effects related to prominence (Saltzman et al., 2008). Studies have also investigated articulatory kinematic patterns during acoustic pauses occurring at prosodic boundaries, which led to a proposal of pause posture (Katsika et al., 2014; Krivokapić et al., 2020). For the more general overview of prosody within AP, see Byrd and Krivokapić (2021) and Krivokapić (2020).

Overall, pitch gestures as well as prosodic gestures were proposed to account for the suprasegmental aspects of speech under the framework of AP, yet there are many areas that call for further exploration. One such example is the mapping of phonological primitives – i.e. pitch gestures – into actual F0 values. The current gestural model of F0 control is thus not only a tool for testing the competing hypotheses of pitch control (*target* vs. *register*) but also contributes to the expansion of the AP research.

4.2 Gestural model of F0 control

4.2.1 Basic mechanisms

The two most crucial components of the gestural model are (i) the normalized F0 tract variable and F0 gestures and (ii) pitch register parameters. I first introduce the part of the model about normalized pitch gestures and tract variable and discuss how they are mapped to actual F0 values through pitch register parameters.

The current study adopts the dynamic field model of movement preparation which was proposed by Erlhagen and Schöner (2002) and was adapted to speech tasks by Tilsen (2007, 2009, 2018). Under this model, each tract variable is associated with a distribution of activity in the intentional planning field. When gestures associated with a tract variable become active, they create a Gaussian distribution of activation in the field, and the activation centroid of the field determines the target value of the tract variable. Namely, gestural targets specify the distributions of input activity in the field, and the tract variable targets arise from integrating these inputs to the fields (Tilsen, 2018). This perspective is different from the standard AP, where the gestural target is a scalar value that specifies an equilibrium or target of the tract variable. Hereafter, I refer to the target of the tract variable as *dynamic* target to distinguish it from *gestural* target, which is the target parameter of the gestures.

As mentioned in Section 4.1.1, F0 is as also considered as a tract variable, which is analogous to other tract variables such as LA, TTCL, or TTCD, and F0 gestures change the equilibrium of the F0 tract variable. When pitch gestures are turned on, they exert Gaussian forces on the planning fields. Likewise, a neutral attractor also exerts forces on the planning fields. The dynamic target of the F0 tract variable is then derived

from the centroid of the summation of two forces – one from the F0 gestures that are active and the other from the neutral attractor. If, however, there are no F0 gestures active in the field, the neutral attractor force will only affect the tract variable target. See Figure 4.3 for an example of the intentional planning field, which shows the dynamic target of the tract variable (black dashed line) calculated from the two forces active in the field – high F0 gesture (red line) and neutral attractor (blue line). The targets of pitch gestures and neutral attractors as well as the dynamic targets of F0 tract variable are all specified in the normalized field coordinates, in the interval from 0 to 1, as shown in Figure 4.3. The current model no longer assumes gesture-specific stiffness; stiffness is instead determined by the activation in the field and a field-specific gain parameter.

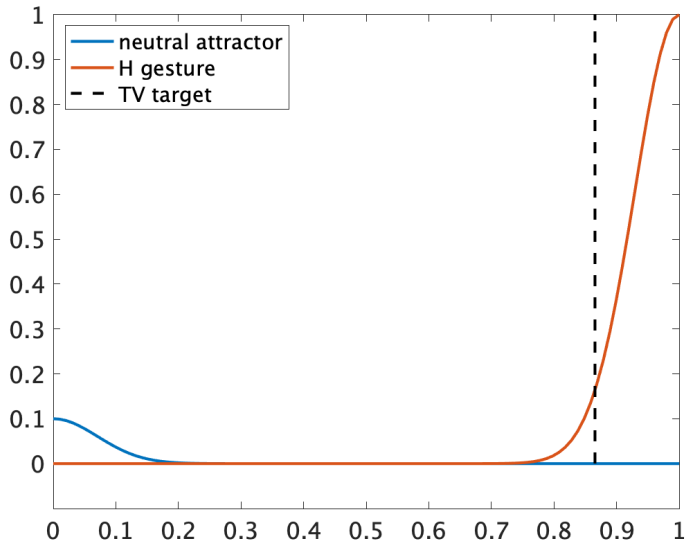

Figure 4.3. An example of the intentional planning field. The planning field coordinates are normalized in the interval of [0-1]. The blue line shows the force of the neutral attractor, and the red line shows the force of the H gesture. The neutral attractor and the pitch gesture create Gaussian forces on the field. In this example, the mode of the H gesture is set as 1, and the mode of the neutral attractor as 0; in both cases, the sigma is 0.1. The amplitude of the neutral attractor is set as 0.1, and the gestural target as 1. The target of the tract variable (black dashed line) is derived from the summation of the two forces active in the field.

The normalized tract variable targets are then mapped to actual F0 values in Hz with pitch register parameters. Specifically, the normalized dynamic target values are first multiplied by the parameter value of the register span, and they are added onto the parameter value of the register floor. Here, I specifically adopt the register span and floor parameters to map the abstract tract variable values into actual F0 values, but other combinations of register parameters (e.g. ceiling and span) are also possible. In the *register*-control variations of the model (i.e. Models 3 and 4 in Table 4.1), pitch register parameters can reset in the middle of the utterance, for instance at a phrasal boundary; this allows gestures with identical target values to be realized as distinct F0 values. Moreover, the register floor parameter decays at a constant rate to model the effect of global declination.

4.2.2 Parameters

The gestural model of F0 control is implemented with the following set of parameters. See Table 4.2, which lists the name and definition of all model parameters. The first set of parameters is defined for each gesture specifically ((i) gesture-specific parameters in Table 4.2); these are gestural targets, onsets, and durations.

In the second set ((ii) gesture-independent parameters), parameter values do not vary by individual gestures, but the same values (either specified or optimized) are applied to all gestures in the tract variable. The first is the mode and the sigma of the neutral attractor (specified in normalized units); in our model, the neutral attractor exerts constant forces to the planning fields. The next is the ramp parameter, which is relevant to the activation ramping and determines how smoothly the gestures are turned on and

off – i.e. it determines the amount of increase and decrease in gesture strength at the edges of its activation intervals. An identical ramp parameter value is used for all gestures. As mentioned above, the stiffness (which specifies how fast the equilibrium is achieved) in the current model is no longer defined for each gesture but is instead determined by the activation in the field and the gain parameter of the field. The model also has a gain parameter for the register floor and span to allow for a smooth transition of register between phrases, in case it varies at a phrasal boundary. The declination parameter specifies the global decay of the register floor, and therefore, is gesture-independent.

The last set of parameters is associated with pitch register. The model specifically adopts register floor and span parameters to model the tonal space of a speaker when producing a given F0 contour. These parameters specify the initial value of the floor and span of the utterance, if register is defined utterance-specifically (Models 1 and 2 in Table 4.1), but the initial value of the phrase if register varies from phrase to phrase (Models 3 and 4 in Table 4.1).

The optimization algorithm searches for the best values of these parameters within a specified set of lower and upper bounds. The bounds of the parameters as well as their initial guesses will be introduced in Section 4.3.2.

Table 4.2. A list of parameters in the gestural model. The parameters are categorized as (i) gesture-specific, (ii) gesture-independent, and (iii) pitch register-related. Each row presents the name of the parameter along with its definition.

name	**definition**
(i) gesture-specific parameters	
target	target of the gesture in normalized units [0-1]
onset	onset of the gesture in seconds (relative to the beginning of the phrase)
duration	duration of the gesture in seconds
(ii) gesture-independent parameters	
neutral attractor mode	mode of the neutral attractor in normalized units [0-1]
neutral attractor sigma	sigma of the neutral attractor in normalized units [0-1]
ramp	activation ramping
gain	gain for the planning field (instead of gesture-specific stiffness)
floor gain	gain for floor
span gain	gain for span
declination	register floor declination in normalized units/s
(iii) pitch register parameters	
floor	initial register floor for a phrase or utterance in Hz
span	initial register span for a phrase or utterance in Hz

4.2.3 Optimization

The parameters of the model were optimized using a global optimization method. The use of the global optimization technique was necessary, as the parameter space was very high dimensional, and our goal was to search for the global minimum of a function that contains multiple local minima.

Among various global optimization methods, pattern search optimization was specifically conducted using the Matlab global optimization toolbox (cf. the motivation for selecting this solver is introduced in Section 4.3.3). The optimization algorithm

searches for a set of parameters that returns the lowest cost between the input F0 contour and the optimized contour; the cost was the root mean squared differences between the two contours.

4.2.4 F0 models and experiments

F0 models were fit to the smoothed and interpolated F0 contours of the subject phrase. Specifically, F0 contours of the seven participants who showed similar accentual patterns in the production experiment (Chapter 3) were used to test the gestural model; see Figure 4.4 for the average time-warped F0 contours of all experimental conditions of one such participant. The data of these participants had one F0 peak, one F0 valley preceding the peak, and another F0 valley following the peak for each subject NP. Thus, one high (H) and one low (L) pitch gesture were posited for each NP.

The composition of H/L gestures of each phrase, however, can be determined in alternative ways, for instance by algorithmically identifying the number of peaks and valleys of each phrase and varying the number of H and L gestures accordingly (see Section 4.5.4.1). Yet, it is important to emphasize that the goal of the modeling is *not* to find a contour that perfectly fits the empirical contour. This is because some of the F0 peaks and valleys in the empirical data may be the results of microprosodic variation or F0 tracking irregularity at voicing transition. For instance, if we were to find a perfect fit for the contours in Figure 4.4, we should not only model the F0 peak and the preceding and following valleys of each NP, but also smaller peaks around phrasal boundaries (vertical dashed lines); in this case, it would be better to posit more H and L gestures rather than a single set of H and L for each NP. Smaller F0 variations can be

meaningful in some cases, but as the first step of testing our gestural F0 model, I focus on the largest-scale F0 variations observed in the empirical F0 trajectories and posit one H and one L gesture at each phrase.

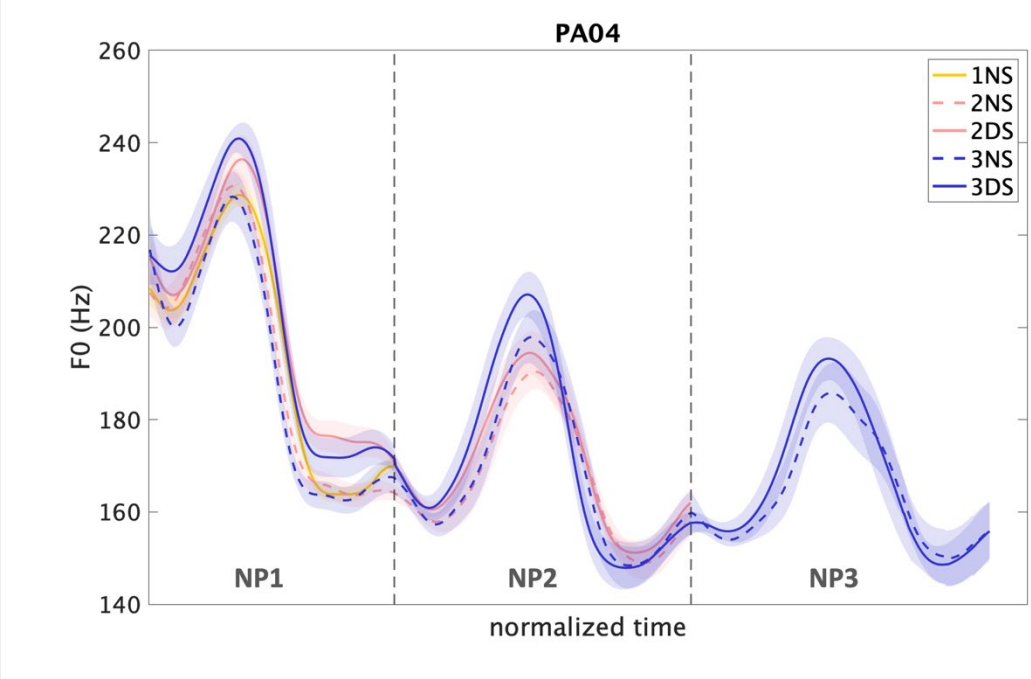

Figure 4.4. Examples of the smoothed and interpolated F0 contours that were time-warped by each subject NP. The F0 contours of seven participants (including this participant) were examined in the modeling. All NPs had an F0 peak and F0 valleys preceding/following the peak, which let us posit one H and one L F0 gesture for each NP. See Figure 3.6 in Chapter 3 for the specific information about the figure and F0 contours of six other participants who showed a similar accentual pattern. Note that the y-axis of Figure 3.6 is F0 that was recentered within each participant; for the modeling, the original F0 values were used.

The main hypotheses of F0 control – i.e. *target* vs. *register*-control – were reflected as constraints on parameters in the model. Specifically, four different F0 models were constructed depending on whether gestural target parameters or register parameters are shared across phrases (defined *by-utterance*) or vary from phrase to phrase (defined *by-phrase*). The four models are listed in Table 4.3 and their schematic

illustrations are presented in Figure 4.5.

In Model 1, the target values of pitch gestures associated with NPs are the same, and a single set of register parameters is introduced to model the subject F0 contour (i.e. H-L/FL-SP). In Model 2, gestural targets of each NP are allowed to have distinct values, although a single set of register parameters is still assumed for the whole subject phrase (i.e. H1-L1-H2-L2/FL-SP). In Model 3, F0 gestures associated with different NPs have the identical target values; yet, the register parameters are specified for each phrase (i.e. H-L/FL1-SP1-FL2-SP2). This allows for a variation of register in the middle of the utterance – e.g. register shift or register span expansion/compression. For simplicity, register parameters are only allowed to vary at a phrasal boundary in the current model, which can be improved in future works. Lastly, in Model 4, gestural targets are varied by phrase, and also, a unique set of register parameters is specified for each NP (i.e. H1-L1-H2-L2/FL1-SP1-FL2-SP2). Thus, Model 4 is the most flexible (i.e. allows maximal variation of model parameters), Model 1 has the least freedom, and Model 2 and Model 3 are similar in terms of complexity, as only one set of parameters (target vs. register) is allowed to vary by phrase.

Linking our conceptual hypotheses of F0 control with the F0 models in Table 4.3, the *target*-control hypothesis is implemented as Model 2, as Model 2 assumes that speakers have different gestural targets for each phrase. On the other hand, the *register*-control hypothesis matches Model 3, as Model 3 assumes that speakers have different representations of tonal space that are specific to each NP, and this phrase-specific register results in F0 variations. Note that Model 4 can be understood as an exemplification of the combination of two hypotheses (*target* and *register*-control

hypotheses), as both gestural targets and register parameters vary from phrase to phrase; see Section 4.5.2 for further discussions.

For a given F0 contour, all four versions of the models were fit to the data, and their performances were examined by comparing their fits. To reiterate the hypotheses and predictions introduced in Section 4.1, the *target*-control hypothesis would predict a lower cost in Model 2 compared to Model 3, whereas the *register*-control hypothesis would predict a lower cost in Model 3 than Model 2. Model 1 is expected to perform the worst, as both gestural targets and register parameters are fixed at an utterance-level. Model 4, on the other hand, is expected to show best performance, as both target and register parameters are tailored to each phrase. However, it is likely that Model 4 is too powerful: since there are so many parameters to optimize and various possible solutions, this may complicate the algorithm, resulting in a failure to find the best fit. It is also possible that other less complicated models (i.e. Models 2 or 3) can provide an equally good fit as Model 4.

Table 4.3. Four different F0 models that are tested in this study. Gestural targets and register parameters can be defined at an utterance-level (by-utt) or at a phrase-level (by-phr): see the first two columns. The third column provides a detailed explanation about the model, including the parameter setting for the data that have two NPs in the subject phrase. H/L: H/L gestural targets, FL: floor, SP: span. The 1 and 2 indicate the specific NP that is associated with the parameters.

target	**register**	
Model 1.		
by-utt	by-utt	The targets of H and L gestures are shared across phrases, and a single set of floor/span parameters is defined for the utterance. (e.g. H-L/FL-SP)
Model 2.		
by-phr	by-utt	The targets of H and L gestures vary from phrase to phrase, but a single set of floor/span parameters is defined for the utterance. (e.g. H1-L1-H2-L2/FL-SP)
Model 3.		
by-utt	by-phr	The targets of H and L gestures are shared across phrases, but a set of floor/span parameters is defined for each phrase. (e.g. H-L/FL1-SP1-FL2-SP2)
Model 4.		
by-phr	by-phr	The targets of H and L gestures vary from phrase to phrase, and a set of floor/span parameters is defined for each phrase. (e.g. H1-L1-H2-L2/FL1-SP1-FL2-SP2)

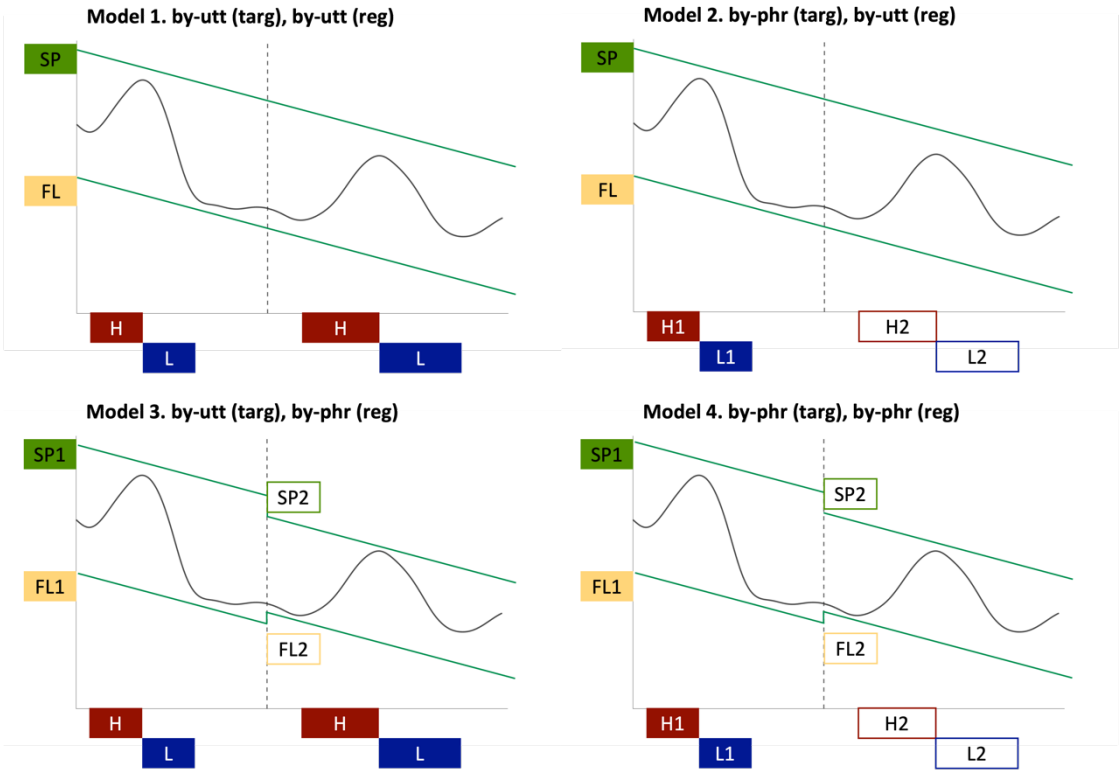

Figure 4.5. Schematic illustrations of four F0 models and the empirical contour. H1/L1 and H2/L2 indicate that distinct targets are specified for each gesture of a phrase. SP1/FL1 and SP2/FL2 likewise indicate the phrase-specific register parameters.

In the implementation of the models, the four models in Table 4.3 were further broken down by whether gestural targets and register parameters are fixed at certain values or whether they are optimized. When gestural targets or register parameters are optimized, the algorithm explores the parameter space and attempts to find the parameter values that best fit the original F0 contour; but when they are fixed, the parameters are not optimized, and the model uses the specified parameter values to generate the output F0 contour.

When gestural targets are fixed, H targets were always fixed at the value of 1 and L targets at 0. It was reasonable to choose these values (rather than some random values), as it was sensible to assume that the H targets are at the top edge of one's tonal

space, and the L targets to be at the bottom edge. Note that the values of 1 and 0 do not mean the top and bottom of the physiological pitch range of a speaker, but the edges of the current pitch register. For the fixed register parameters, the values were chosen speaker-specifically, based on all F0 values of a given speaker. I reasoned that the fixed register should characterize a speaker's usual tonal space as closely as possible, and the more reliable estimation of register can be obtained if all F0 productions of a given speaker are considered. Thus, F0 values of all trials of a given speaker were first collected, and then their minimum F0 value was used as the fixed floor parameter, and their range multiplied by 1.25 was used as the fixed span parameter. The reason why the range was multiplied by 1.25 (rather than just using the range) was due to the forces of neutral attractors. Since neutral attractors constantly exerted forces on the planning fields, extra room of F0 was needed to properly map gestural targets to actual F0 values. Several multipliers – i.e. 1.1, 1.2, 1.25, 1.5, 2 – were tested on the sample of 10 trials from two speakers, and the model with the multiplier 1.25 showed the best performance.

Overall, a combination of four models in Table 4.3 and the fixed vs. optimized distinction altogether resulted in a total of 16 different models. See Table 4.4 for the full list of all F0 models. In the table, notice that the model 2a, 2b, and 4a, 4b are missing, but instead they are marked as 1a*, 1b*, and 3a, 3b*. This is because in implementation, the models 2a, 2b, 4a, 4b are identical to 1a, 1b, 3a, 3b, respectively: when gestural targets are fixed, they are always fixed at 0 (L target) or 1 (H target), which makes the by-phrase and by-utterance distinction meaningless. Thus, these redundant models were not tested in the current study (they are colored in grey in Table 4.4), which leaves a total of 12 different models. It is generally expected that the more parameters are

optimized, the better optimized result is obtained; thus, within each of Model 1, 2, 3, and 4, the "d" variant is expected to show the best performance.

Table 4.4. A full list of F0 models tested in the current study. A combination of four models in Table 4.3 and the fixed vs. register distinction results in 16 different models. The rows that are colored in grey are not tested in the current study.

	by-utterance vs. by-phrase		**fixed vs. optimized**	
	target	**register**	**target**	**register**
1a	by-utterance	by-utterance	fixed	fixed
1b			fixed	optimized
1c			optimized	fixed
1d			optimized	optimized
1a*	by-phrase	by-utterance	fixed	fixed
1b*			fixed	optimized
2c			optimized	fixed
2d			optimized	optimized
3a	by-utterance	by-phrase	fixed	fixed
3b			fixed	optimized
3c			optimized	fixed
3d			optimized	optimized
3a*	by-phrase	by-phrase	fixed	fixed
3b*			fixed	optimized
4c			optimized	fixed
4d			optimized	optimized

Two types of modeling experiments were conducted: the first is the *speaker*-level modeling, and the second is the *trial*-level modeling. In the speaker-level modeling, the parameters of all 12 models in Table 4.4 were fit to the average time-warped F0 contours generated for each participant and experimental condition – for example, the F0 contours in Figure 4.4. Based on the results of the speaker-level modeling, selected models were tested in the trial-level modeling; here, F0 contours of all trials of all participants were evaluated by the models.

4.3 Methods

4.3.1 Data

The smoothed and interpolated F0 trajectories of the subject phrase were modelled. The data from seven participants who showed similar accentual patterns were specifically examined; see Figure 4.4 and Figure 3.6 in Chapter 3 for their F0 patterns. F0 values were not recentered, as the focus of the analysis was not on the across-speaker comparisons of F0 contours, but on comparing the model fits for each individual contour.

In the speaker-level modeling, F0 models were tested on the linearly time-warped F0 contours that were generated for each participant and experimental condition. F0 contours from each NP were warped to the median length of that NP, and their average was subject to modeling. A total of 35 F0 contours (7 participants x 5 conditions) were examined by 12 F0 models presented in Table 4.4.

In the trial-level modeling, F0 models were tested on the F0 trajectory of each individual trial. A total of 1680 trials, which remained after a series of outlier removals (Section 3.2.3, 3.2.4), were evaluated by the selected models.

4.3.2 Parameter setting and inequality constraints

In this section, I introduce the initial guesses (*b0*) (i.e. the starting parameter value for the optimization) and the lower (*lb*) and upper bounds (*ub*) of each parameter and the motivation for selecting those values. Setting reasonable bounds and initial guesses is crucial to obtain good model fits. Table 4.5 lists all parameters with their bounds and

initial values; the table also indicates whether each parameter was optimized (free to vary) or fixed. Figure 4.6 provides an example of a time-warped F0 contour that had two NPs in the subject phrase (blue lines) along with an F0 contour and pitch register that were generated with the initial parameter values of each model (yellow lines).

Table 4.5. Initial guesses and lower and upper bounds of each model parameter. The first column shows the name of the parameter, and the third column (b0) shows the initial guess. The second (lb) and fourth (ub) columns represent the lower and upper bounds. The final column (fixed/free) indicates whether each parameter was fixed or optimized in the model.

name	***lb***	***b0***	***ub***	**fixed/free**
(i) gesture-specific parameters				
target (H)	0.5	1	1	fixed/free
target (L)	0	0	0.5	fixed/free
onset (relative to phrase onset)	0	phrase dur / # of gestures	phrase dur	free
duration	0.2	middle value between 0-phrase dur	phrase dur	free
(ii) gesture-independent parameters				
neutral attractor mode		0		fixed
neutral attractor sigma		0.2		fixed
ramp	0.02	middle value between *lb-ub*	0.1	free
gain	3	middle value between *lb-ub*	200	free
floor gain		500		fixed
span gain		500		fixed
declination	-F0 range / dur	0	0	free
(iii) pitch register parameters				
floor	F0 min	*lb*	F0 max	fixed/free
span	F0 range	*lb*	F0 range × 2	fixed/free

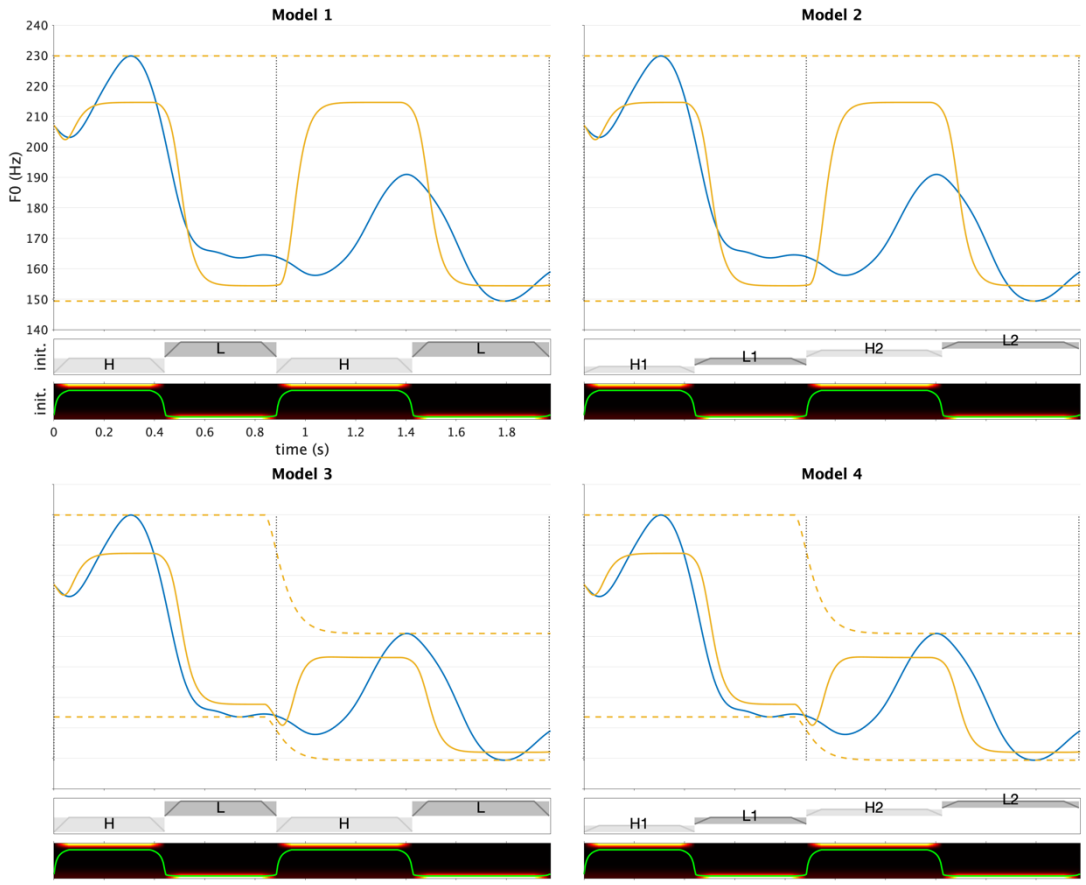

Figure 4.6. An empirical F0 contour (identical with the contour in Figure 4.5) and the initial model-generated F0 contour. For each figure, the blue line in the first panel shows the empirical F0 contour that had two NPs in the subject phrase; the vertical dotted line marks the end of the first NP. The yellow solid line shows the F0 contour generated by the model with the initial guess of the parameters; the yellow dashed lines illustrate the initial register. The second and the third panels show the initial gestural score and intentional planning field, respectively. Note that the gestural scores simply show the order and timing of the gestures; the height of the gesture does not represent its target value.

Gestural targets could either be fixed or optimized. See Table 4.4 for the fixed vs. optimized setting of each model. As mentioned, if gestural targets were fixed, H targets were fixed at 1, and L targets were fixed at 0. If they were optimized, the *lb/ub* of H target were set as 0.5 and 1, and the *lb/ub* of L target were set as 0 and 0.5. These bounds

were selected, as the targets of H and L gestures should be located within the upper and lower half of tonal space, respectively.

The onset and duration of an F0 gesture were always optimized. Since the onset parameter was defined relative to the onset of the phrase that the gesture was associated with, the *lb* was set as 0 (i.e. the gesture starts at the beginning of the phrase), while the *ub* was set as the phrase duration (i.e. the latest possible starting point is the end of the phrase). The initial guess of this parameter was dependent on the location of a given gesture within the phrase: if it was the first gesture of the phrase, *b0* was set as 0, while in other cases, *b0* was calculated as the phrase duration divided by the number of gestures in the phrase. In our experiments, since each NP was always assumed to be composed of one H and one L gesture, the initial guess of the H gesture was always 0, while it was phrase duration/2 for the L gesture. See the initial starting point of each gesture in the gestural score (second panel) of Figure 4.6.

Regarding gestural durations, *ub* was set as the phrase duration, and *b0* was set as the middle value between 0 and *ub*; the gestural score in Figure 4.6 shows the initial gestural duration. For the *lb*, instead of setting it as 0, I set it as 0.2, which meant that each gesture should at least be 200 ms long. If we allowed gestural durations to be an extremely small number that is close to 0, gestures may not exist in the optimized results although we posited them in the gestural score. Moreover, an extremely short gesture could be used to model F0 variations that we are not really interested in – for example, microprosodic variations or F0 irregularity. For this reason, the *lb* of the duration was set to be above 0, specifically 200 ms.

The parameters associated with the neutral attractors were always fixed, as there

was no reason to vary them by individual F0 contours. The target of the neutral attractor was set as 0 – meaning that it has a mode at the register floor – and the sigma was set as 0.2.

The parameters for activation ramping and field gain were always optimized, yet the gain for the floor and span parameters were fixed. The *lb/ub* for the ramp parameter was 0.02 and 0.1, and the middle value was used as the initial guess. For the field gain, the *lb/ub* was 3 and 200, and the initial guess was the middle value between the *lb* and *ub*. The gains of the floor and span were always 500. The choice of these numbers was rather arbitrary – I did not want the values of these parameters to be too small nor too large; yet, in some cases, the bounds were selected after testing on the a few empirical contours.

Regarding declination, the *lb* was set as -F0range/dur, and the *ub* was set as 0. The *lb* assumes the maximal declination given the empirical data[7] (i.e. F0 declines gradually and constantly throughout the utterance), and the *ub* assumes no global declination. The initial guess was same as *ub*; in Figure 4.6, the floor does not show any declination.

The register parameters could be fixed or optimized depending on the model setup. As mentioned above, if they were fixed, F0 minimum of all F0 values of a given speaker was used as the floor parameter, and F0 range × 1.25 (due to the neutral attractor) was used as the span parameter.

If they were optimized, the *lb* of floor and span were set as F0 minimum and range,

[7] I assume that, in our empirical data, speakers produced F0 within pitch register, without hitting or going below the bottom of the register. Therefore, the *lb* of the declination parameter is the maximal declination that can be found within the register without hitting the floor.

respectively, of an utterance or a phrase. Specifically, when the model assumed utterance-specific register (i.e. Models 1, 2 in Table 4.3), the *lb* of the floor and span were the minimum and range of F0 of the whole subject phrase; on the other hand, when the model assumed phrase-specific register (i.e. Models 3, 4 in Table 4.3), the *lb* of floor and span were the minimum and range of F0 of each phrase. These were the reasonable *lb* values, as there is no reason to assume that the optimized floor and span go below the minimum and range observed in the empirical F0 contour.

On the other hand, F0 maximum and F0 range × 2 were used for the *ub* of the floor and span parameters, respectively. The floor should specify the bottom of one's tonal space, and therefore, it could never exceed the maximum F0 value of the empirical contour. For span, doubling the empirical F0 range was an arbitrary yet a logical choice, as one's usual tonal space would not be larger than twice the range of the empirical contour, especially given that the empirical contour was produced at a normal pitch range (i.e. not with an extremely small voice). The initial guess for the floor and span parameters was set as *lb*, which is demonstrated in Figure 4.6; the initial floor and ceiling (floor + span) are at the minimum and minimum + range of the whole subject phrase in Models 1 and 2, while they are at the minimum and minimum + range for each subject phrase in Models 3 and 4.

Given this parameter setting, the information on the complexity of each model (i.e. the number of parameters) is added to Table 4.4 and introduced again in Table 4.6. The last two columns indicate the number of total parameters and free (i.e. optimized) parameters that were needed to model an F0 contour with three NPs in the subject phrase. The number of total parameters increased from Model 1 to 4 – i.e. Model 1: 23,

Model 2 & 3: 27, Model 4: 31, as more parameters were needed when gestural targets and register vary by phrase. Within each model, the "a" variant had the fewest optimized parameters, and the "d" variant had the greatest number of free parameters.

Table 4.6. F0 models tested in the current study with information on their complexity. This table copies Table 4.4 with additional information on the number of total and free parameters needed to model F0 contours with three noun phrases.

	by-utt vs. by-phr		**fixed vs. optimized**		**complexity**	
	target	**register**	**target**	**register**	**total #**	**free #**
1a	by-utt	by-utt	fixed	fixed	23	15
1b			fixed	optimized	23	17
1c			optimized	fixed	23	17
1d			optimized	optimized	23	19
1a*	by-phr	by-utt	fixed	fixed		
1b*			fixed	optimized		
2c			optimized	fixed	27	21
2d			optimized	optimized	27	23
3a	by-utt	by-phr	fixed	fixed	27	15
3b			fixed	optimized	27	21
3c			optimized	fixed	27	17
3d			optimized	optimized	27	23
3a*	by-phr	by-phr	fixed	fixed		
3b*			fixed	optimized		
4c			optimized	fixed	31	21
4d			optimized	optimized	31	27

Besides parameter bounds, two inequality constraints were imposed to obtain better model fits. The first was that the gestures in the same phrase cannot overlap for more than a certain value, which we set here as 100ms. This constraint was imposed to prevent too much overlap between the H and L gestures. The second constraint was that the duration of a gesture cannot extend beyond the duration of the phrase that the gesture is associated with. Since gestures are supposed to model F0 peaks and valleys found in each NP, it was reasonable to constrain them within their associated NP (at least for the

English data that are used in this study).

4.3.3 Optimization testing

4.3.3.1 Global optimization solvers

For the current study, using a flexible global optimization technique was necessary for the reasons mentioned in Section 4.2.3. The general goal of the global optimization is to search for the global minimum of a function that contains multiple local minima, and this is in line with the goal of the current experiment. Before committing to a particular global optimization method, I tested four different optimization solvers – i.e. pattern search, particle swarm, genetic algorithm, and surrogate – to find the best solver for our dataset. These solvers were specifically chosen, as they allow bounds for each parameter, enable parallel computing (which would save time to fit models on all data), and work on both smooth and non-smooth problems. All optimizations were conducted in Matlab using the global optimization toolbox. The detailed explanations for the four optimization solvers can be found in the following websites, and Table 4.7 provides a summary.

https://www.mathworks.com/help/gads/

https://www.mathworks.com/products/global-optimization.html

https://www.youtube.com/watch?v=4wgI3-RQqTY

Table 4.7. Global optimization solvers tested in the current study. The first column shows the solver name, and the second column briefly explains how it finds solutions.

solver	**description**
pattern search	- an approach that searches a set of points ("mesh") generated through pattern vectors around a current point - expands or contracts if a solution is not found
particle swarm	- a collection of particles moves throughout a region, and the algorithm evaluates the function at each particle - particles have velocity and are affected by other particles in the swarm
genetic algorithm	- starts with an initial generation of candidate solutions - subsequent generations evolve toward an optimal solution through selection, crossover, and mutation
surrogate	- an approach that creates and optimizes a "surrogate" of the function that is usually expensive to evaluate - searches randomly to explore and adaptively to refine

Each solver has a set of optimization options which allow fine-tuning of the solver performance. For instance, most optimization solvers have the *MaxIterations* option which specifies the maximum number of solver iterations and the *MaxTime* option which specifies the total time (in seconds) that are allowed for the solver to find solutions. Among various options, those that are relevant to the initial setting of the search (that makes the initial search more fine-grained) were varied to identify the best set of optimization solver and the option parameter value. Table 4.8 introduces the name of the optimization option that was varied in the test, as well as the default value of that parameter, and the range of the parameter values that was tested. For the test values (the last column of Table 4.8), I specifically chose the default value, default × 10, and default × 100, except for the genetic algorithm, in which a large parameter value resulted in a model failure. It should be noted that the current solver/parameter testing is not exhaustive, since only a single optimization option was varied, and the number of test

values was limited; I thus acknowledge that the optimization solver and the parameter setting determined from this test are not necessarily most optimal for the current study.

Table 4.8. Optimization options tested in this study, which adjust the setting of the initial search. The first column shows the solver, and the second column introduces the option. The third and the fourth columns show the default parameter value for the given option and the range of values that is tested. Note that in some cases the default value can vary by the number of variables in the function that is evaluated, and the values presented here are those relevant to the current function.

solver	**optimization option**	**default value**	**test values**
pattern search	*InitialMeshSize*	1.0	1.0, 10, 100
particle swarm	*SwarmSize*	100	100, 1000, 10000
genetic algorithm	*PopulationSize*	200	200, 300, 400
surrogate	*MinSurrogatePoints*	2 x num of vars	default, default × 10, default × 100

A combination of four models and a range of parameters (Table 4.8) was tested on 10 trials selected from the speaker-level experiment data. Model 3b was specifically used for the solver/parameter testing (cf. this could have created a bias for a better performance of this model; see Section 4.5.2 for further discussions). The two inequality constraints (i.e. gestures cannot overlap over certain amount or extend beyond phrases) mentioned in Section 4.3.2 were also imposed, except when using particle swarm, which did not allow inequality constraints. In all cases, the *MaxTime* option was set as 60 min. The average cost (i.e. the difference between the modeled contour and the input contour) and the average time it took for the solver to find solutions were examined and compared across the solver/parameter sets.

Table 4.9. Optimization solver and option testing results. The top table shows average cost (RMSE, in Hz), and the bottom table shows average time (in seconds) it took for the solver/option to find solutions. Test value 1 is the default parameter value. For pattern search, particle swarm, and surrogate, test values 2 and 3 are default×10 and default×100, respectively; for genetic algorithm, each of them represents 300 and 400.

	test value 1 (default)	**test value 2**	**test value 3**
average cost (in Hz)			
pattern search	1.98	1.99	2.01
particle swarm	2.55	2.41	1.99
genetic algorithm	2.06	2.11	2.15
surrogate	4.86	4.92	7.59
average time (in seconds)			
pattern search	140.3	142.0	139.3
particle swarm	89.2	235.4	1875.2
genetic algorithm	395.5	627.3	905.6
surrogate	23.3	25.9	21.6

The results from solver/option parameter testing are presented in Table 4.9. In terms of the average cost (top part of Table 4.9), all test values of pattern search and the test value 3 of particle swarm showed lower costs (i.e. average cost less than or around 2 Hz). Yet, among these cases, the average time it took for the solvers to find solutions (bottom part of Table 4.9) was much shorter when using the pattern search method than the particle swarm method. This overall led us to choose the pattern search solver for the current modeling experiments. Since the three test values did not show a large difference in terms of both average costs and time, a more detailed option testing for the pattern search solver was conducted, which is introduced in the next section.

4.3.3.2 Pattern search solver

Further investigations on the options of the pattern search solver were conducted to find the optimal setting for the current experiments. In this test, *InitialMeshSize* which

was varied in the previous test as well as *MaxIterations* and *MaxFunctionEvaluations* options were examined. This allowed variations in the fineness of the initial search setting (*InitialMeshSize*) and the maximum numbers of solver iterations (*MaxIterations*) and function evaluations (*MaxFunctionEvaluations*), thus increasing a possibility for the solver to find the best solution.

Three test values were chosen for each of the three optimization options – the default value, default × 10, default × 100, which resulted in a total of 27 sets of options and parameter values. These sets were tested on a single trial that had three NPs in the subject phrase. The results found the lowest cost for the given trial in the following setting: *InitialMeshSize* set as default × 100, *MaxIterations* as default × 10, and *MaxFunctionEvaluations* as default × 100. Thus, the pattern search solver with this set of option parameter values was used in all F0 modeling of the current study. As mentioned above, this solver/setting may not be optimal for the entire data, since the parameter tests were only conducted on a single F0 contour, and only selected options and parameter values were tested.

4.3.4 Data analysis

In the speaker-level experiment, 12 different F0 models presented in Table 4.4 were fit to 35 average time-warped F0 contours, and their performances were examined. To ensure that all F0 models worked as intended, cases which potentially suggest a modeling error were algorithmically identified. The first was the case in which the optimized F0 contour simply shows a straight line – i.e. no prominent F0 peaks and valleys. To identify this case, a linear model was fit to input F0 data, and the RMSE

between the linear fit and the input contour was calculated. When the RMSE calculated with the linear fit and the RMSE of the model fit showed little difference – i.e. RMSE linear fit / RMSE model fit < 1, that modeling result was identified as error. The second was the case in which the optimized F0 contour goes outside the optimized floor and span. Instances of both cases were not observed in the results of the speaker-level experiment. For further confirmation, I plotted and sanity checked all model fits (optimized F0 contours and floor/span) in the speaker-level experiment. To compare the performance of 12 models, the average cost was calculated over 35 trials for each model. The cost was the root mean squared differences between the optimized F0 contour and the input contour.

In the trial-level experiment, four different F0 models – i.e. Models 1d, 2d, 3d, 4d – were fit to a total of 1680 individual F0 contours. These four models were specifically selected, as they allowed comparisons among Models 1, 2, 3, and 4 (these comparisons were only possible among the "c" and "d" variants), and the cost of the "d" variants was lowest within each model (which will be introduced in Section 4.4). As in the speaker-level experiment, cases where the modeled F0 contour showed a straight line or cases with the optimized contour above/below the optimized register were identified. No linear fits were observed, but there were 238 cases (3.54%) out of 6720 model fittings (1600 trials x 4 models), where the optimized contour exceeded the optimized register. These cases were excluded from subsequent analyses.

The performance the four models was also compared with RMSE. Rather than comparing the average cost between the models (as in speaker-level experiment), the costs of the four models were compared *within* each trial, and the differences in model

costs were examined over the whole dataset. The within-trial model comparison was considered appropriate, as the effects of variation in individual F0 contours could be minimized on evaluating the model performance (this is analogous to the motivation for using a paired two-sample t-test as opposed to an unpaired test). To statistically show that the cost significantly differed by model type for each trial, the Friedman test, which is a non-parametric statistical test similar to the repeated measures ANOVA, was conducted. The non-parametric test was used, as it is likely that the model costs are not normally distributed. If model type was found to be a significant factor, pairwise comparisons between the four models were conducted via the paired Wilcoxon signed-rank test using the Bonferroni correction method.

4.4 Results

Overall, the gestural model of F0 control, which was newly developed and tested in this dissertation, fit the empirical data with relatively small differences. The performances of Model 2 and Model 3 were of particular interest, as they implemented the *target*-control and *register*-control hypotheses, respectively; in Model 2, gestural targets varied from phrase to phrase with utterance-specific register, while in Model 3, register parameters were specified for each phrase with utterance-specific gestural targets. It was found that in both speaker-level and trial-level experiments, Model 3 showed a lower cost than Model 2. This result provides evidence that speakers are mainly controlling pitch register to produce F0 variations.

Section 4.4.1 discusses the general performance of the gestural F0 model by

presenting the minimum and average model costs and sample model fits. In Sections 4.4.2 and 4.4.3, the results of the speaker-level and trial-level modeling experiments are introduced. In presenting the results below, the model that exhibited lower cost is the model that showed better performance, as it means that the optimized model fit showed smaller differences from the input F0 contour.

4.4.1 General model performance

The gestural model of F0 control showed overall good performance when it was tested on the data collected from the production experiment. The smallest root mean squared error (RMSE) that was found in the speaker-level experiment was 0.26 Hz, and it was 0.15 Hz in the trial-level experiment. The average cost calculated over the mean time-warped F0 contours was 1.98 Hz, and the average calculated over the contours of all trials was 2.57 Hz.

Figure 4.7 presents sample modeling results. Each figure shows the empirical average time-warped F0 contour (blue solid line) and the modeled F0 contour (orange solid line) and register (orange dashed lines), with its cost in the parentheses in the title. The figures also present the optimized gestural scores and intentional planning fields. These three cases had the lowest cost among trials with one, two, and three NPs in the subject phrase for this speaker, PA02. The first figure is the result from Model 1d, and the rest of the figures are the results from Model 4d. The figures demonstrate that our F0 model captured the major F0 variations (i.e. F0 peak, preceding and following valleys) fairly well, although it could not model secondary F0 peaks around phrasal boundaries, presumably due to the insufficient number of F0 gestures.

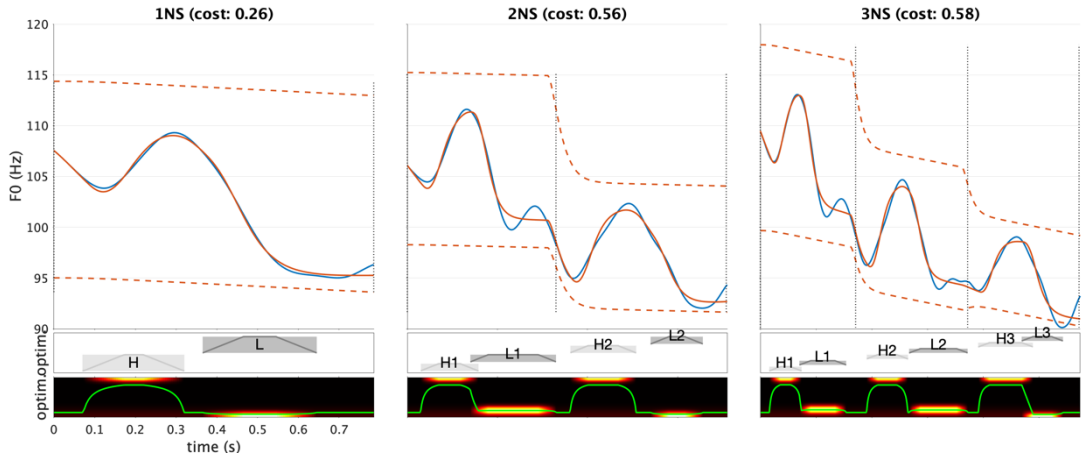

Figure 4.7. Examples of model fitting. The figures show the modeling results of the average time-warped F0 contours of PA02. Each figure is composed of three panels: the first panel shows the input F0 contour (blue solid line), modeled F0 contour (orange solid line), and modeled register (orange dashed lines). The second and third panels show the optimized gestural score and intentional planning field. The gestural score presents the order and timing of the optimized gestures, but not the values of the gestural targets. The centroid of the activation forces is plotted in the planning fields. The first figure shows the average time-warped contour of condition 1NS that was modeled with Model 1d. The second and third figures show the contours of 2NS and 3NS, respectively, that were modeled with Model 4d. In Model 4d (second and third figures), both gestural targets and register parameters were defined phrase-specifically; the labels H1/H2 and L1/L2 in the gestural score were used to indicate distinct gestural targets for each NP; the register also varied around a phrasal boundary, which is represented as a vertical dotted line in the first panel.

In addition to the relatively low costs of the models, the other evidence that the implementation of our F0 models was sensible is that the model performance did not significantly differ in trials that had a single NP in the subject phrase (trials in condition 1NS). Figure 4.8 compares the costs of four models – Model 1d, 2d, 3d, 4d – within the trials of the same experimental condition – 1NS, 2DS, 2NS, 3DS, 3NS. Unlike trials with two or three NPs, trials with a single NP did not show a large difference in costs by model type. This is expected given that the difference in the model specification (1d vs. 2d vs. 3d vs. 4d) comes from whether the gestural targets and register parameters are

defined at an utterance-level or at a phrase-level; thus, for a single NP subject which is a phrase and also an utterance, the results should not differ crucially by model type. For this reason, the modeling results of the trials with a single NP subject were excluded from the subsequent model comparisons. Note also that the cost increased as more NPs were added into the subject phrase (1NS < 2DS, 2NS < 3DS, 3NS).

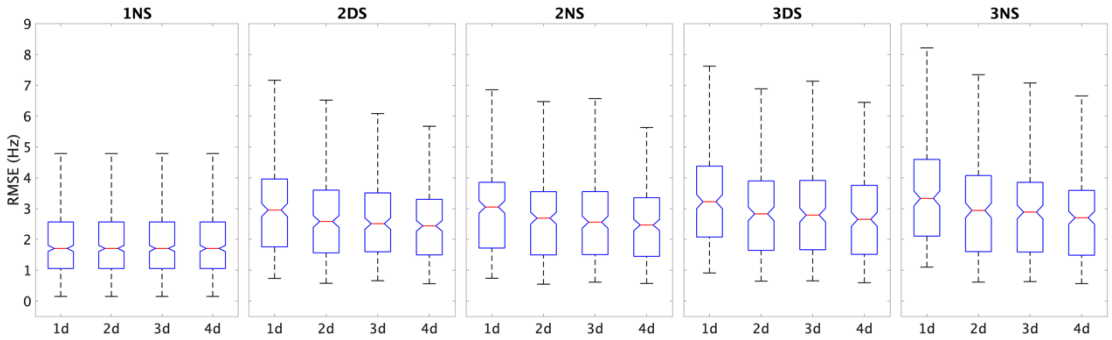

Figure 4.8. Comparison of four models (Models 1d, 2d, 3d, 4d) in each experimental condition. In each panel, the distribution of model costs (RMSE) is presented by model type. The red line within each box shows the median, and the edges of the box show the 25-75th percentile. The title of the panel indicates the experimental condition.

4.4.2 Speaker-level experiment

12 different F0 models presented in Table 4.4 were fit to the average time-warped F0 contours of each participant and experimental condition, and the variants of Models 3 and 4 provided good fits of the data. Specifically, among 12 models, the average cost was lowest in Model 4d (1.59 Hz), which was followed by Model 3d (1.64 Hz), and then Model 3b (1.67 Hz). Figure 4.9 and Table 4.10 show the average costs of each model calculated over 28 time-warped F0 contours that had two or three subject NPs. (cf. Each model was originally fit to 35 average F0 contours, but the results of 1N trials (7 contours) were excluded from the analyses.)

The important finding is that the variants of Model 3 showed good performance, almost as similar as Model 4. It was expected that Model 4 would produce a good fit, as both gestural targets and register parameters were defined phrase-specifically, and thus, the model had more freedom to reflect the F0 properties of each phrase. Yet, when only one set of parameters (targets vs. register) was allowed to vary – thus, between Models 2 and 3, the model in which the register was specified for each NP showed better performance. This result provides support for the *register*-control hypothesis.

Another interesting point is that Model 3b, in which the gestural targets were fixed at 1 (H target) and 0 (L target) and only register parameters were optimized, showed a good model fit. This may be because the solver/parameter testing was conducted specifically with this model (Section 4.3.3), so the solver option was more fine-tuned to this model. Yet, alternatively, it may suggest that the speakers have their H and L targets fixed at the top and bottom edge of the register, and the control of tonal space results in the variations of F0.

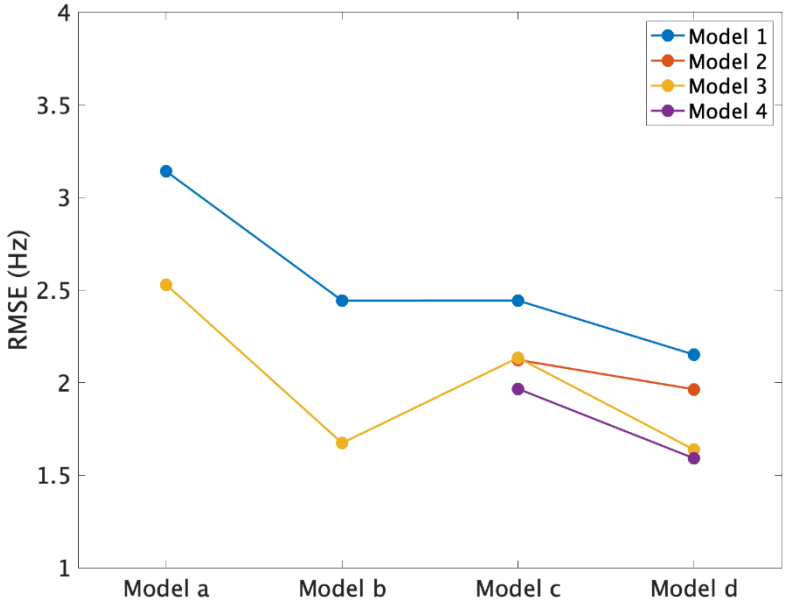

Figure 4.9. Mean optimized costs of 12 F0 models. The blue line represents Model 1, where both gestural targets and register parameters were defined at an utterance-level. The orange line shows Model 2, where gestural targets varied by phrase, and the yellow line shows Model 3, where register parameters varied across phrases. Model 4, in which both targets and register were defined phrase-specifically, is shown as a purple line. Models a-d varied by whether the parameters of interest (gestural targets and register) were optimized or fixed. As shown in Table 4.4, Models 2a, 2b, 4a, and 4b were not examined.

Table 4.10. Mean optimized costs of 12 F0 models calculated over 28 time-warped data. This table copies Table 4.4 and Table 4.6 with information on the average cost added in the final column.

	by-utt vs. by-phr		fixed vs. optimized		complexity		average cost
	target	**register**	**target**	**register**	**total #**	**free #**	**(Hz)**
1a	by-utt	by-utt	fixed	fixed	23	15	3.14
1b			fixed	optimized	23	17	2.44
1c			optimized	fixed	23	17	2.44
1d			optimized	optimized	23	19	2.15
1a*	by-phr	by-utt	fixed	fixed			
1b*			fixed	optimized			
2c			optimized	fixed	27	21	2.12
2d			optimized	optimized	27	23	1.96
3a	by-utt	by-phr	fixed	fixed	27	15	2.53
3b			fixed	optimized	27	21	1.67
3c			optimized	fixed	27	17	2.13
3d			optimized	optimized	27	23	1.64
3a*	by-phr	by-phr	fixed	fixed			
3b*			fixed	optimized			
4c			optimized	fixed	31	21	1.97
4d			optimized	optimized	31	27	1.59

It was also found that among the four variants of the model (i.e. Models a, b, c, d), the "d" variant showed the best performance, and the "a" variant showed the worst performance, when the four variants were compared within each model (i.e. within Models 1 and 3). For those with just the "c" and "d" variants (i.e. within Models 2 and 4), the cost of Model d was lower than that of Model c. This result confirmed our hypothesis that the modeled F0 contour exhibits a smaller difference from the input contour, when parameters are allowed to vary rather than fixed. This is because if the parameters are optimized, they could be more tailored to the specific characteristics of each F0 contour. As the "d" variant had the lowest cost within all models, Models 1d, 2d, 3d, and 4d were specifically tested in the trial-level experiment.

The findings from the speaker-level experiment can be summarized as follows: (i) the models that had phrase-specific register (especially when the register is optimized) produced good model fits (i.e. Models 3b, 3d, 4d); (ii) the models showed better performance when more parameters were optimized (i.e. the "d" variant within each model). These findings are confirmed graphically in Figure 4.10, which shows the optimization results of 12 models that were tested on the same F0 contour (i.e. average time-warped F0 contour of trials in condition 2DS of PA04). The blue (input contour) and the orange (modeled contour) lines are very close to each other in Models 3b, 3d, and 4d (the bottom three figures in the right column); except for the small F0 rise observed at the end of NP2, the input and the modeled contours are almost identical in the rest of the utterance. Moreover, the difference between the blue and the orange lines was smallest in the "d" variant within each model.

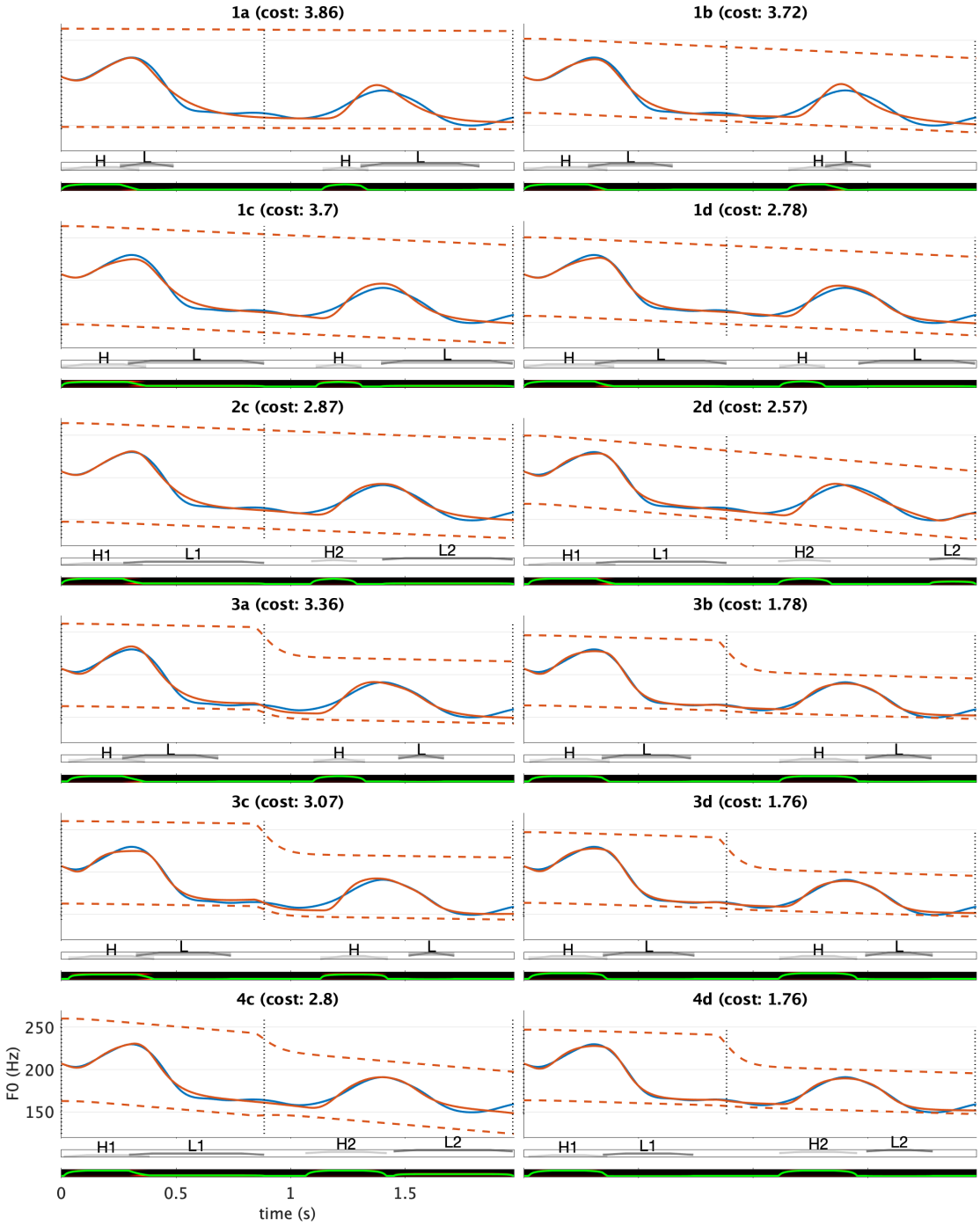

Figure 4.10. Comparison of optimization results of 12 F0 models. The input F0 contour is identical in all cases (i.e. average F0 contour of trials in 2DS of PA04). Each figure shows the input contour (blue) and the optimized contour/register (orange), along with the optimized gestural score and the planning field. The title shows the model type and cost (Hz).

4.4.3 Trial-level experiment

We have observed in the previous section that the models with phrase-specific register (i.e. Models 3 and 4) provided good fits of the data, which supports the *register*-control hypothesis. The goal of the trial-level experiment is to find out whether this holds in the modeling of individual F0 contours. The selected models – i.e. Model 1d, 2d, 3d, and 4d, in which both gestural targets and register parameters were optimized, were fit to the F0 contours of individual trials, and their costs were examined. This time, model comparison was conducted within each trial.

The results of the trial-level experiment were similar to the speaker-level experiment, as the models with phrase-specific register showed good performance. Figure 4.11 compares model fitting results of the two sample trials – one is from PA01 (top row) and the other is from PA04 (bottom row). In both cases, Models 3d and 4d showed good fits of the data, exhibiting less differences between the blue line (input contour) and the orange line (optimized contour).

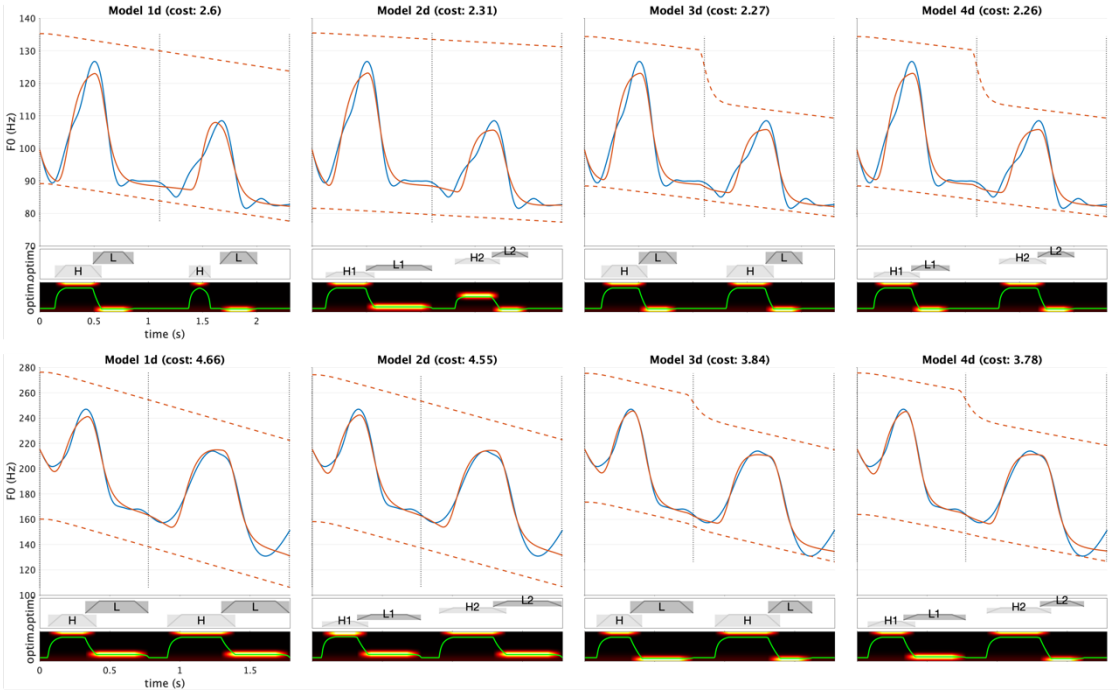

Figure 4.11. Comparisons of optimization results of four selected models. The figures in the top row show model fits of a single trial from PA01, and the figures in the bottom row show model fits of a trial of PA04.

The similar result was observed when the within-trial cost differences were aggregated and examined. Figure 4.12 presents the distributions of cost differences between the models. From the first three boxplots (Δ 1d-2d, Δ 1d-3d, Δ 1d-4d), it was found that Model 1d had in general higher cost than the other models – i.e. the medians of these boxplots were all above 0; moreover, Model 1d showed the largest difference with Model 4d – i.e. the median of Δ 1d-4d was highest among the three boxplots. Comparison between the fifth (Δ 2d-4d) and the sixth (Δ 3d-4d) boxplots further showed that the costs of Models 2d and 3d were in general higher than the cost of Model 4d – i.e. the median of the two boxplots were above 0. A larger difference with Model 4d, however, was found at Model 2d, which suggests that Model 3d performed better than Model 2d. This was also confirmed in the fourth boxplot (Δ 2d-3d) – i.e. the median

was close to but slightly above 0. This result suggests that Model 3d provided a better fit of the data than Model 2d on average, although the difference between the two models was not large, and there were trials where the reverse was true (i.e. cases where Δ 2d-3d is below 0).

Overall, the observations from Figure 4.12 can be summarized as follows. First, the cost of Model 1d was higher than the other models, suggesting its poor performance. Second, the cost of Model 4d was lower than the other models, suggesting its good performance. Note, however, that it was not always the case that Model 4d showed better fits than the other models – for instance, there were cases where Δ 2d-4d was below 0 (i.e. Model 2d had a lower cost than Model 4d). This suggests that Model 4d, although in general successfully finds the best fit compared to the other models, it may fail to do so; this shows that the model with too many free parameters can indeed be overly powerful. Lastly, the cost of Model 3d was slightly lower than that of Model 2d. This result is important as it provides supporting evidence for the *register*-control hypothesis over the *target*-control hypothesis. However, the difference between these two models was not very large; this suggests that pitch register can be understood as the main control parameter in general, though there is some evidence that the opposite might be true.

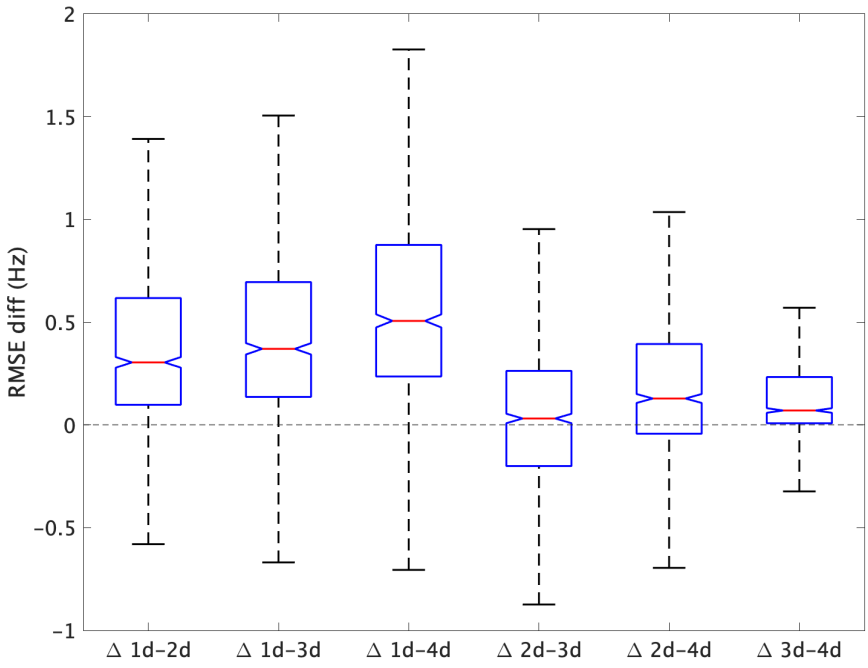

Figure 4.12. Differences in model costs calculated within each trial and aggregated over dataset. The y-axis shows RMSE differences; the x-axis indicates which models are compared. The horizontal dashed line at 0 shows that the model costs were same in the two models that were compared.

To further show that our model comparison results are statistically significant, the Friedman test was conducted with model costs of each trial (not the cost differences) as the dependent variable, and model type (i.e. Model 1d, 2d, 3d, 4d) as the independent variable. The purpose of the statistical test was to examine whether the distributions of the costs crucially differ by the four paired groups. The results found a significant effect of model type on the costs (X^2 (3) = 1414, $p < 0.001$). Post-hoc tests were then conducted to find out which models showed significant differences. The results from the pairwise Wilcoxon signed-rank tests are presented in Table 4.11; they showed that all four models significantly differed from each other.

Table 4.11. Statistical results from the pairwise Wilcoxon signed-rank tests. The first two columns indicate which models are compared. All groups had n = 1001, which is the total number of 2DS, 2NS, 3DS, 3NS trials without any modeling errors. The third column shows test statistic that is used to compute p-values. The final column indicates the significance level.

group 1	group 2	test statistic	significance level
1d	2d	462230	$p < 0.001$
1d	3d	458960	$p < 0.001$
1d	4d	478102	$p < 0.001$
2d	3d	275472	$p < 0.05$
2d	4d	371151	$p < 0.001$
3d	4d	433012	$p < 0.001$

4.5 Discussion

4.5.1 Overall evaluation of the model

In this chapter, an F0 model based on the framework of Articulatory Phonology was proposed and evaluated with the empirical data. The model showed good performance, as the difference between the input F0 contour and the model-generated contour was in general small. Indeed, most of the differences arose from features of the contours that we did not aim to capture – in particular, the secondary small F0 peaks that occurred at phrase boundaries (see Figure 4.7, Figure 4.10, Figure 4.11).

The average root mean squared differences between the empirical and optimized contours calculated over all time-warped contours (speaker-level experiment) was 1.98 Hz, while the average calculated over individual contours (trial-level experiment) was 2.57 Hz. For comparison, Kochanski et al. (2013) fit the soft-template model to the Mandarin Chinese corpus, and the RMSE they obtained was 13 Hz. In Yuan (2004),

who modeled F0 contours of statements and questions in Mandarin Chinese with the soft-template model, the lowest RMSE was 9.4 Hz. Compared to these studies, the current F0 model seemed to fit empirical data fairly well, though it is not necessarily possible to make direct comparisons between the results of these studies and the current one, as F0 contours that were modeled differed significantly.

The other evidence that showed that the gestural model worked well as intended came from the model comparisons on the trials that had a single NP in the subject phrase. In the experiments, four models (i.e. Model 1, 2, 3, 4) were fit to the empirical contours, and their performances were examined. These models differed by whether variations of F0 peaks and valleys observed in the data are modeled by phrase-specific vs. utterance-specific gestural targets and register parameters. For instance, different F0 peaks and valleys could arise by allowing only gestural targets to vary across phrases (Model 2), or only register parameters to vary (Model 3), or both targets and register to vary (Model 4). These models, however, should not differ when they are fit to the F0 contours of the trials that had only a single NP in the subject phrase. This is because this single NP is a part of the subject phrase, but at the same time, it is the subject phrase (i.e. utterance in our term) itself; this makes the distinction of by-phrase or by-utterance target/register meaningless. Moreover, it is unclear whether the F0 peaks and valleys observed in this NP come from the variations of F0 targets or register, as both can basically produce the same result.

As expected, the costs of the four models did not show a large difference for those trials that had a single NP subject. This was particularly evident in Figure 4.8, in which the distributions of the costs of Models 1d, 2d, 3d, and 4d were almost identical in the

1NS trials, unlike the trials in other conditions, which showed large differences. These results altogether show that the gestural model of F0 control was successful in fitting our empirical data and further suggest that the model is a valid and powerful tool to examine our hypotheses of F0 control.

4.5.2 Model comparisons I: by-utterance vs. by-phrase

The main finding of the modeling experiments is that the F0 model in which the register parameters were allowed to vary by phrase while gestural targets were constant (Model 3) outperformed the model in which the gestural targets varied by phrase with constant register (Model 2). This result suggests that it is pitch register that speakers are controlling to produce variations in F0.

In the speaker-level experiment, out of 12 different models, Models 4d, 3d, and 3b had relatively lower costs. These are the models that had phrase-specific register, and the register parameters were optimized. A particularly interesting finding is that Model 3b produced good fits almost as similar as those of Model 3d. The difference between the two models was that the gestural targets were fixed at 1 (H target) and 0 (L target) in Model 3b, while they were optimized in Model 3d. This may suggest that speakers have a fixed target, which is at the edges of the register; yet, the result may have simply arisen because the solver/option testing was conducted specifically with Model 3b.

The better performance of Model 3 over Model 2 was also demonstrated in the trial-level experiment. In this experiment, four selected models (Models 1d, 2d, 3d, 4d) were fit to the F0 contours of individual trials and their performances were compared within each trial. It was found that Model 4d (by-phr target & register) showed the best

performance, which was followed by Model 3d (by-utt target & by-phr register), then Model 2d (by-phr target & by-utt register), and Model 1d (by-utt target & register). The poor performance of Model 1d was expected, as it was the least specific model. In Model 1d, the optimization algorithm must find a single set of gestural targets and register parameters to model the contours of the entire subject phrase, and thus, the performance of this model cannot be very good. In a similar sense, the good performance of Model 4d was also expected, as the algorithm was allowed to find the best set of target/register parameters for each NP and thus had more freedom to reflect the specific properties of F0 contours of each phrase. It is, however, possible that Model 4d is overly powerful; since there may be multiple solutions for a given F0 contour, the model may not be able to find the most optimal one. We have indeed observed this possibility from cases where Model 2d had a lower cost than Model 4d in Figure 4.12.

The key analysis, therefore, was the comparison between Models 2d and 3d, in which either gestural targets or register parameters were allowed to vary by phrase and the other parameter was set constant. These models were also directly relevant to the main hypotheses of this dissertation – i.e. *target* vs. *register*-control hypothesis. The results found that Model 3d fit the empirical data better than Model 2d, although the difference between the two models was not very large. Yet, even if the performance differences across models did not lead to strong inferences, as a proof of concept, this study has demonstrated that the register control is a powerful control mechanism which can produce variations in F0. What this result means is that when speakers vary one component of F0 – i.e. gestural targets vs. pitch register, it is likely that they vary pitch register. In other words, for the utterance that is composed of multiple phrases with

various peaks and valleys, speakers might have an invariant representation of high and low targets throughout the utterance (rather than several distinct high and low targets for each phrase), and the actual realization of the highs and lows depends on the control of tonal space; i.e. speakers have one set of cognitive representations of high and low throughout the utterance, but they control F0 space to realize the abstract representations into different surface F0 peaks and valleys.

It is, however, important to note that our finding does not provide conclusive evidence for the control of register in F0 production. There are a number of objections that may be valid, specifically regarding model design and its implementation; these will be introduced in Section 4.5.4 along with some suggestions for future research. Still, our modeling experiments are valuable in that they allowed explicit comparisons of F0 control hypotheses through parameter optimization.

Based on our modeling results, one can argue that speakers are in fact controlling both register and targets, as Model 4, in which both targets and register parameters were defined phrase-specifically, showed the lowest cost. This is indeed a possibility, but at the same time, the result can simply arise due to the nature of optimization – i.e. when more parameters are optimized, the model is more likely to return better fits. Comparing the number of free parameters between Models 1d, 2d, 3d, and 4d (Table 4.6), Model 1d had the fewest free parameters (19), which was followed by Models 2d and 3d (23), and then Model 4d (27).

The model with more free parameters is likely to produce better fits of the data, yet this does not hold true across the board, specifically regardless of model type. Figure 4.13 presents the costs of the average time-warped F0 contours that had three NPs in the

subject phrase with respect to the number of free parameters in the model; model types (Model 1-4) are distinguished by colors, yet no distinctions are made regarding whether targets or register parameters were fixed vs. optimized (Model a-d). The figure shows a tendency in which the optimized cost decreases as more parameters are optimized, yet we could still observe the importance of model type. For instance, in cases where the models had 15 or 17 free parameters, data from Model 3 (yellow dots) had a relatively lower cost than the data from Model 1 (blue dots). Similar patterns were observed in cases where the models had 21 or 23 free parameters; the cost was lower in models with phrase-specific register (Model 3) or phrase-specific register and target (Model 4).

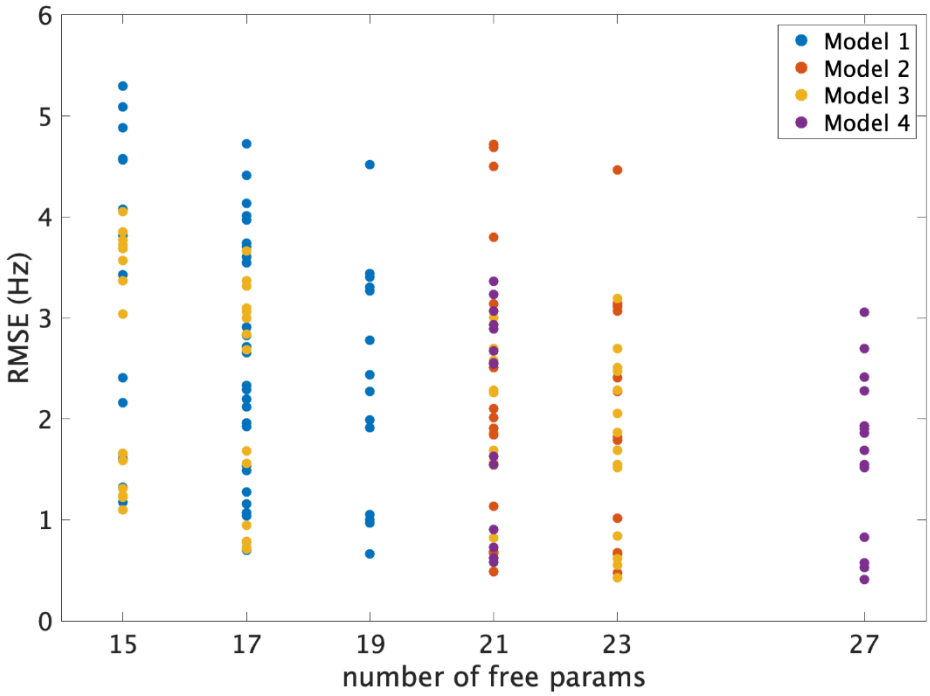

Figure 4.13. Distributions of model costs according to the number of free parameters in the model. Each datapoint shows the model cost of the mean time-warped F0 contours of trials with three subject NPs. Datapoints from Model 1 are presented in blue, Model 2 in orange, Model 3 in yellow, and Model 4 in purple. No distinction is made on whether gestural targets and register parameters are fixed or optimized.

These observations suggest that we cannot attribute the good performance of Model 4 entirely to the number of free parameters, which makes the argument that speakers control both gestural targets and pitch register more persuasive. If one can develop an algorithm where model cost is calculated with consideration of the number of free parameters in the model – e.g. penalize the model with more free parameters, it would provide better insights on the performance of Model 4, whether it is due to the nature of the optimization algorithm or indeed reflects the speakers' F0 control mechanism.

4.5.3 Model comparisons II: fixed vs. optimized

Within Models 1, 2, 3, and 4, the model that optimized both gestural targets and pitch register (the "d" variant) outperformed the models that optimized either one or the model that used only fixed values. In the speaker-level experiment, the average cost of the "a" variant (fixed target & register) was in general higher than the "b" (fixed target & optimized register) and "c" variants (optimized target & fixed register), which were higher than the "d" variant (optimized targets & register). This is because if parameters are optimized, the algorithm can find the best parameter values that are specific to each individual F0 contour. This model is likely to provide a better fit than the models which use parameter values that are fixed for a given speaker.

In that sense, to make the fixed vs. optimized comparisons sensible, it is important to come up with good parameter values for the fixed targets and register. When gestural targets were fixed, they were fixed at 1 (H target) and 0 (L target). This was a logical choice, as it is more reasonable to assume that the targets are fixed at the edges of the

register rather than at some random points; yet, these values could be tested in future work. The fixed values of the register parameters can also be adjusted. To determine the fixed values of register parameters in a way that captures the usual F0 space of a speaker as closely as possible, I collected F0 values of all trials for a given speaker and used F0 minimum as the fixed register floor and F0 range × 1.25 as the fixed register span. The choice of 1.25 as a multiplier was not completely random, as it was selected after testing five different multipliers (i.e. 1.1, 1.2, 1.25, 1.5, 2) on the F0 contours of 10 trials from two speakers. It is yet possible that different multipliers could provide better fits for the entire data. Therefore, further systematic tests on the fixed parameter values should be conducted to improve the performance of some variants of the models and thus to better compare the fixed vs. optimized settings.

4.5.4 Limitations and future research

The newly developed gestural model fit our empirical data with relatively high precision, yet there is certainly room for improvement in various aspects of the model. In this section, I present the limitations of the current F0 model and experiments and offer some future directions. I first discuss improvements that can be made on the model assumptions on the empirical data, specifically about gestural composition and location of register shift. I then discuss some limitations on model parameters and optimization algorithm. Lastly, I present how the current model can contribute to the research of Articulatory Phonology in general.

4.5.4.1 Assumptions on the empirical data

In the current model, based on the empirical F0 pattern which had one F0 peak and F0 valleys preceding and following the peak, one H and one L F0 gesture were posited for each NP. With this gestural composition, the models were able to capture the major F0 variations that I was interested in; yet, as shown in the figures that display the sample fits (Figure 4.7, Figure 4.10, and Figure 4.11), the model could not account for all F0 peaks and valleys observed in the data. This could be improved if one can develop an algorithm which posits pitch gestures according to the specific properties of F0 contours – for instance, algorithmically identify the numbers of F0 peaks and valleys and posit H and L gestures accordingly. The use of this algorithm, however, would then require decisions about which F0 peaks and valleys to posit as gestural, since we do not want to model all sorts of F0 variations that are even microprosodic.

The other assumption that needs further exploration is where in the utterance the register should vary. In this study, I have limited register changes (if allowed) to occur only at a phrasal boundary. This is because F0 range is known to reset at intermediate or intonational phrase boundaries in English (e.g. Beckman & Pierrehumbert, 1986), and these prosodic units are likely to arise at the end of each subject NP given its syntactic structure. Also, it is better to constrain the region where the register can vary rather than allowing it to occur at any time in the utterance, as the model may be too powerful in the latter case. Yet, it is possible that speakers prepare for the register shift before reaching the boundary, not necessarily *at* the boundary. One way of dealing with this possibility is to optimize the timepoint that the register shifts can occur, within a specified range of periods around a phrasal boundary – for example, allowing register

to vary at any time after the F0 peak of a given NP and before the peak of the following NP. This would allow F0 models to find a better timepoint for register shift in a relaxed but still constrained manner.

4.5.4.2 Model parameters

The initial guesses and the upper and lower bounds of model parameters can be further tested. For some parameters – e.g. the mode and sigma of the neutral attractor, ramp, and gain, the choice of initial values and bounds was rather arbitrary, and thus, setting different values for these parameters may provide better model fits. In particular, the mode of the neutral attractor was always fixed at 0, which means that the neutral attractor force has a mode at the register floor. However, if we assume that speakers do not make use of the full tonal space that is available at a given time in an utterance, using different values (e.g. 0.25, 0.5) may be more adequate – i.e. setting the resting position (i.e. neutral attractor) to be a little beyond the register floor.

The declination parameter also calls for some further exploration. In the current version of the model, the declination parameter lowered the register floor at a constant rate. Considering that the register span was not affected by the declination parameter, our model assumes that the effect of declination is identical on both register floor and ceiling. It is, however, possible to obtain more accurate model fits when the two register parameters are differently affected by declination – i.e. have a separate declination parameter for each of the floor and span. Moreover, the way that the declination parameter affects the register floor can be altered. Shih (2000), for instance, proposed an exponential declination model, based on the data which showed a faster declination rate at the beginning of the utterance. Fujisaki (1983), on the other hand, modeled the

register floor to rise at the beginning and asymptotically decline over the course of the utterance. It is also possible that the model can perform well even without any explicit declination parameter, as assumed in Liberman and Pierrehumbert (1984).

4.5.4.3 Optimization algorithm

There are also some points to make regarding the optimization algorithm. First, we cannot guarantee that the optimization algorithm always finds the global minimum. Second, it is possible that there may be a set of equivalent or nearly equivalent optimal solutions – i.e. a hyperplane in the parameter space, which indeed may be the case for Model 4. These points suggest that the results we have obtained may not be necessarily the best fit of the input F0 contour, which can be a potential challenge to our conclusion.

The optimization solver and solver options can also be systematically tested. The current study used the pattern search solver with adjustments in the optimization options that are relevant to the setting of the initial search (i.e. *InitialMeshSize, MaxIterations, MaxFunctionEvaluations*). Although the solver and the option parameter values were selected after they were tested on some of our empirical data, they were not evaluated on the entire dataset, which leaves a possibility that the current optimization setting is not optimal for our data. However, the space of hyperparameters for solvers and for models, is far larger than it can be investigated, and thus, it will always be necessary to attempt to make informed decisions, as I have done here.

4.5.4.4 Articulatory Phonology research

The introduction of the gestural F0 model with dynamic register suggests some future directions for the research of Articulatory Phonology in general. In particular, the mapping between F0 primitives (i.e. F0 gestures) and actual F0 values could be better

understood if the coordination between F0 and constriction gestures is also considered. Previous studies have mostly focused on the relative timing between the constriction gestures and the pitch gestures (cf. Section 4.1.1); their coordination, however, may also account for some F0 variations that are induced by vocal tract movements – for instance, the intrinsic F0 of a vowel (i.e. high vowels have a higher F0) or F0 variations of a vowel following voiceless/voiced consonants.

Moreover, we can incorporate modulation gestures such as π-gesture or μ-gesture into the gestural F0 model. For instance, when these modulation gestures are active, are they affecting the targets of H and L pitch gestures or the register parameters or both? Also, how do they influence gestural targets and register – e.g. do they increase gestural targets and broaden register span?

Besides these topics, the novel mechanism of dynamic register can also be applied to other parts of the AP/TD framework. Given the resemblance of tonal space and vowel space as pointed out by Ladd (1992), the concept of dynamic register can be extended to constriction gestures, where the realization of oral gestures is governed by the space of the vocal tract that is available at a given time in an utterance. Further investigations of these questions would enrich not only the gestural model itself, but also expand the research of Articulatory Phonology in general.

CHAPTER 5
CONCLUSION

This dissertation aimed to examine the question of how speakers control F0. I proposed two alternative ways in which the control of F0 can be conceptualized – *target*-control vs. *register*-control, and these hypotheses were evaluated through a production experiment (Chapter 3) and computational modeling (Chapter 4). This chapter provides a summary of the findings from the experiment and the modeling study along with their implications and proposes some future directions.

Section 5.1 summarizes the findings of the experiment, mainly focusing on the speakers' pre-planned and adaptive F0 control. Section 5.2 presents the results from the investigations of the F0 control mechanism; the analyses of F0 control in Chapter 3 and the modeling study of Chapter 4 are summarized. Section 5.3 illustrates some possible future directions of this research, and Section 5.4 concludes this dissertation.

5.1 Pre-planned and adaptive F0 control

Two aspects of F0 control were examined in the experiment: the speakers' (i) pre-planned and (ii) adaptive F0 control with respect to sentence length. Specifically, the purpose of the experiment was to find out whether speakers control F0 (i) according to the utterance length that was presented at the beginning of the trial and (ii) in response to the unanticipated changes in the length.

The results found evidence for both pre-planned/initial and adaptive F0 control. In

particular, participants started from a higher F0 peak and valley and a wider F0 range when they were producing longer sentences. This shows that participants were sensitive to the initially planned sentence length and made F0 plan according to that information. Regarding the adaptive F0 control, participants lowered F0 peaks from the first to the second NP to a lesser extent, when they encountered delayed stimuli. Besides these main findings, the data also showed that participants further lowered a phrase-final F0 valley to mark the end of the subject phrase, and they lengthened phrases (mainly the right edge of the phrase) to plan for the upcoming part of the utterance during production.

A methodological contribution of this study is that I developed a novel experiment paradigm in which the algorithm detected utterance initiation, and the visual stimuli that cued the parts of the utterance were presented once that initiation was detected (rather than at the beginning of the trial). In this case, participants had to adaptively adjust to the changes in the length and content of the sentence and incorporate the delayed stimuli into their ongoing utterance. This perturbation paradigm allowed us to find out whether and how speakers respond to any changes in the stimuli that are made online and identify which F0 parameter is specifically controlled.

The results of this study revealed two important aspects of speech production. One is that the speakers are sensitive to the information of the sentence that they are going to produce and make a sentence plan according to that information; moreover, after they start production, they constantly monitor the sentence (or the environment), and if any changes occur, they respond to them in almost real time. The other important implication is that there is a strong tendency of speakers in which they want their F0 to stay *within* the pitch register. That is, speakers want to secure a sufficient F0 space in order not to

hit the bottom of their F0 range before reaching the end of the utterance. This led them to start from a higher F0 peak or with a wider F0 space when they had to produce longer sentences, and furthermore, reduce F0 peaks to a lesser extent when they saw delayed stimuli.

The current study also made some valuable contributions to the relevant literature. As presented in Section 2.2.3, the literature has been inconclusive as to whether speakers vary utterance-initial F0 peak according to sentence length. The present study supports the claim that speakers do in fact adjust F0 according to the expected length of the utterance. I also demonstrated that it is not just the F0 peak that is influenced by length, but other F0 parameters (i.e. valley, range) are also affected. Furthermore, no previous studies so far have examined whether speakers have an ability to adapt to the changes in the length and content of the utterance; speakers' responses on perturbations in the auditory feedback have mostly been examined (Section 2.2.4). The current experiment in this sense provided evidence that the speakers can adapt to the perturbations that are more fundamental to the structure and meaning of the sentence.

Some of the challenges of the present experiment are the following. First, in this study, trials with potential disfluencies were algorithmically identified based on word and silence interval durations (Section 3.2.3), yet we do not have concrete evidence whether these trials indeed contained any speech errors or hesitations. A potentially better way of identifying disfluencies would be to conduct a perception experiment. With the human-labelled data (and possibly comparing them with the outliers identified algorithmically), we can further examine what contributes to the listeners' perception of disfluencies. Second, further analyses on the F0 trajectories themselves – not the F0

measures derived from them – can be conducted. For instance, F0 trajectories can be analyzed using a generalized additive mixed model (GAMM), which will show whether F0 contours differ by conditions, and in which point in the utterance they start to diverge. In addition, machine learning methods such as clustering or classification can be applied to further identify differences in F0 contours. Lastly, the experiment could have been run on more participants. With more subjects, we could have observed more consistent F0 patterns (not just the major pattern that was analyzed) or obtain more data that exhibit the major F0 pattern, which will result in a greater statistical power.

5.2 F0 control: pitch targets vs. pitch register

The rest of the dissertation examined what it is that speakers control – *pitch targets* vs. *pitch register* – to produce the observed F0 variations. Two forms of investigations were conducted on the data collected from the experiment: the first set of the analyses examined surface F0 measures (F0 peaks/valleys/ranges), specifically their variance and correlation and the condition-prediction regression models in which those measures were used as the predictors; the second set of the analyses focused on the whole F0 trajectories and compared different versions of computational models in terms of how well they fit the empirical F0 contours. The other important difference between the two sets of the analyses was that the pitch targets and register were derived from the surface F0 measures in the first set, such that the F0 peaks were considered as the estimates of H pitch targets, F0 valleys to be the L targets, and F0 ranges to be the register span; in the modeling study, however, the targets and register were optimized given the input F0

contour.

The results from both sets of the analyses supported the control of pitch register. In the first set of the analyses, the variance and correlation measures suggested that F0 peaks and valleys are *not* independently controlled but are correlated. The comparison of the NS vs. DS condition-prediction models yet showed mixed results. In the second set, the dynamical model in which the pitch register varied across phrases outperformed the model in which the targets varied, although the difference between the two models was not very large. Notably, the fact that the two distinct types of analyses all pointed to the *register*-control provides a strong support for such hypothesis.

What this finding suggests is that the speakers have tonal targets that are invariant for a given utterance, yet they control tonal space in which the targets are located to produce various F0 peaks and valleys. This means that speakers have an identical target value for all peaks and another target value for all valleys for a given utterance (the target values are likely to be normalized given the register), and they vary pitch register, where they have options to expand, compress, or shift upwards/downwards, to realize the identical abstract representations into different surface peaks and valleys.

In Chapter 4, F0 control was modelled mainly through the targets of pitch gestures and register parameters. The main feature of the current dynamical model was that the targets of the F0 primitives (F0 gestures) were normalized given the register in the range from 0 to 1, and they were mapped to actual F0 values through the register parameters. Specifically, the model used two register parameters – i.e. floor and span, out of three parameters – i.e. floor, span, and ceiling. In Section 2.2, we have seen that the previous studies assumed the control of register floor and ceiling to model downstep effects,

while they assumed the control of register floor and span (or ceiling and span) for the declination effects. See (a) and (b) of Figure 5.1, which provides schematic illustrations of how register changes in case of downstep and declination in the previous studies. The combination of these effects (assuming that declination is separate from downstep) would look like (c), where the ceiling and floor are lowered (downstep), and at the same time, the span is decreased (declination). The current F0 model aimed to capture this pattern of register changes through variations of register floor and span. Namely, the declination effect was modelled through the lowering of the floor parameter, and by allowing variations in the floor and span parameters, the model could generate F0 peaks and valleys that are downstepped either with a constant or reduced tonal space. (cf. The constant register would be the case where the control of H and L targets is symmetric, and the varying register would be the case when it is asymmetric.)

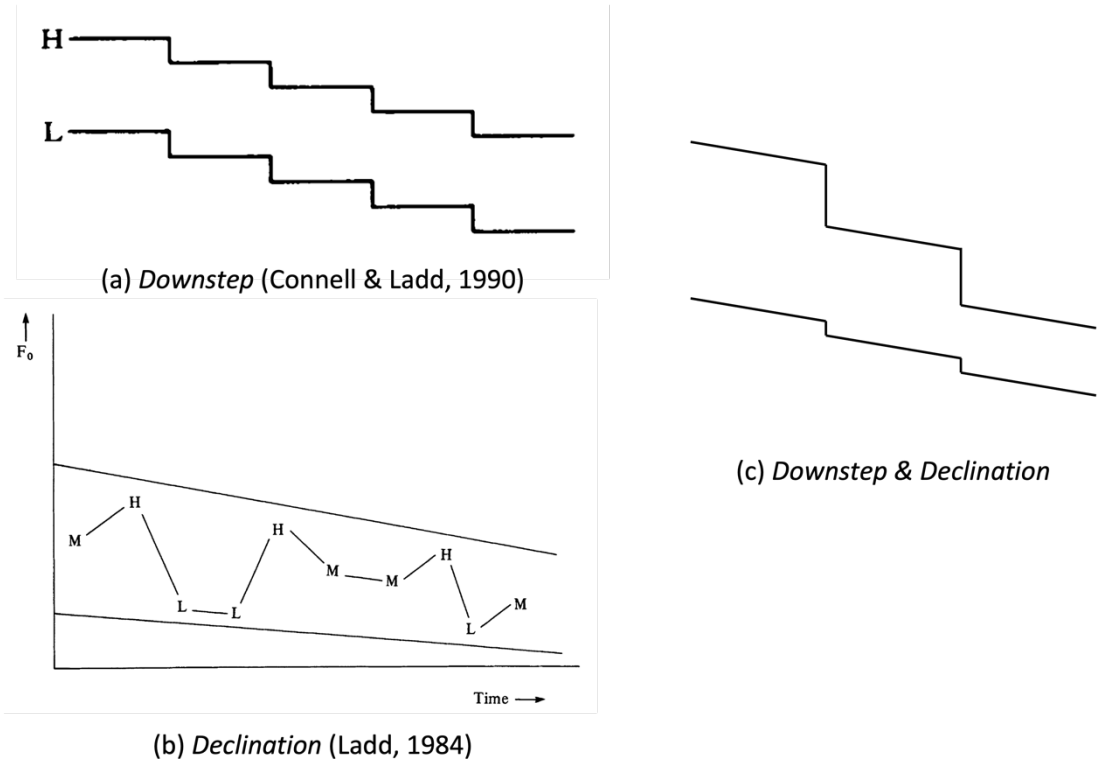

Figure 5.1. Schematic illustrations of downstep and declination effects. (a) and (b) show variations of register in cases of downstep and declination proposed in the previous studies. I predict that the combination of the two effects would look like (c).

This model implementation, however, has some limitations. First, it cannot model the declination that has different shapes – for example, an exponential form of floor lowering as proposed in Shih (2000) or the phrase curve which had a slight initial rise and the asymptotic decline in Fujisaki (1983). Second, the variation of register span may generate cases where the register is indeed expanded over the course of the utterance, which is unlikely to occur in natural intonation without any emphasis/focus. One can, however, prevent this case by imposing an additional constraint on span variation, for instance, that the span of the current NP cannot exceed that of the preceding NP.

Overall, by using register floor and span parameters, the current model could fit

the empirical F0 trajectories with relatively high precision. It is expected that any two combinations of register parameters (e.g. use ceiling/span instead of floor/span) would necessarily produce similar results. Moreover, I believe there would be no cases where we need all three register parameters to model F0 contours, especially sacrificing model complexity (i.e. increasing the number of parameters), mainly because one parameter is simply derived from the other two parameters (e.g. ceiling = floor + span). Yet, these points are speculative and should be empirically tested. It is possible that different sets of register parameters would be more appropriate for some specific languages.

5.3 Future directions

There are several possible future directions of this dissertation. The first is that a more systematic investigation of inter-participant variations can be conducted. The by-participant analysis was briefly presented in Section 3.4.3, yet detailed examinations would demonstrate how participants differ in their extent of controlling F0 according to the sentence length manipulations and which F0 variable they are most likely to use to reflect the length information. Further analyses on the data of participants who showed unique F0 patterns can also be conducted, especially with more data from additional participants. Moreover, it would be interesting to find out whether all speakers control pitch *register* in F0 production, or the choice of *target* vs. *register* differs by speakers.

The second is the analyses on the duration outliers. Due to the novelty of the experiment design, in which the part of the utterance was delayed, it was expected that the participants would produce disfluencies such as hesitations or speech errors. Thus,

before analyzing data, I compared the durations of each word and between-word silence intervals (when present) to identify trials with potential disfluencies and excluded them from subsequent analyses. The duration outliers can, however, be a valuable dataset to understand the speakers' speech planning and processing mechanism. For instance, in cases where disfluencies are detected, where in the utterance are they found? If the disfluencies are observed in sentences with delayed stimuli, are they found at the end of the initially presented NP or within that NP? Also, how is F0 controlled in case of disfluencies? When speakers repair their errors, would they start at an F0 where they left off or rather reset and start from a higher F0? Answers to these questions will better inform us on the speakers' mechanism of speech production and F0 control, which would complement our observations on the trials without any production errors.

Another potential area of future research is whether the idea of *register*-control can be extended to tone languages. In this dissertation, the two alternative hypotheses of F0 control were tested on the intonation language, and it is questionable whether the control of register can also be applied to other types of languages. I expect that the idea of *register*-control would also work for tone languages and account for the empirical phenomena such as tone terracing. In addition, for languages that have more than two tones, the *register*-control hypothesis would provide a better characterization of the speakers' F0 control; rather than arguing that speakers compute individual tonal targets considering the preceding targets which have more than two varieties (i.e. *target*-control hypothesis), it is easier to assume that they have fixed abstract representations for each tone and control tonal space for the surface realizations. Yet, a more careful account is needed when modeling non-automatic downstep, in which the downstep trigger is

absent in the surface tonal sequence.

Lastly, the possibility of target *and* register control could be further examined. In both speaker-level and trial-level experiments of Chapter 4, Model 4 showed the best model fits. I argued that this result was found because there were more free parameters, as both gestural targets and register parameters were fine-tuned to each noun phrase (i.e. they were both defined phrase-specifically); but alternatively, it may in fact suggest that the speakers control both targets and register in F0 production. One way of testing these possibilities is to incorporate ways of penalizing models for their complexity into the comparisons; for example, Model 4 should be penalized more than Models 2 or 3. If Model 4 still outperforms the other models even with this algorithm, it would provide a stronger support for the control of both targets and register. If this is the case, we can further examine whether the speakers' choice of target vs. register control (whether they control only targets, only register, or both target and register) differs by context or by individual speakers. However, such penalization would not be straightforward for the sorts of models that are examined here (as compared with, for example, linear regression models).

5.4 Concluding remarks

Overall, this dissertation examined which hypothesis – i.e. *target* vs. *register*-control – better explains the speakers' cognitive control system of F0. Over the past decades, researchers have made extensive efforts to describe variations observed in the empirical F0 trajectories, to find out which factor/context induces such variations, and

to formally describe F0 contours; yet, the literature crucially lacked investigations of the underlying cognitive mechanism that drives those F0 variations. Presumably, this was because we did not have an appropriate method to measure pitch targets and register (in fact, we can never directly examine these parameters) or derive them from empirical contours.

In that sense, this dissertation constitutes a first step towards understanding the speakers' abstract control of F0. By developing a novel experiment paradigm, we could investigate not only the speakers' pre-planned, initial F0 control but also their adaptive, online F0 control. Moreover, this study examined F0 control parameter by estimating pitch targets and register through the surface F0 measures and through the optimization algorithm in the dynamical gestural models. In sum, this study overall found support for the *register*-control hypothesis. This suggests that the F0 variations may arise from the speakers' control of tonal space, which maps invariant targets of pitch primitives into different F0 peaks and valleys.

APPENDIX

*Table A 1. Results of the mixed-effects linear regressions on the F0 values of the landmarks and rises/falls within each subject NP and F0 differences across NPs. The columns show the dependent variable tested for each regression, and the rows show the conditions of trials included in the model and the effects examined. For example, in the first set of regression, trials in 2NS, 2DS, 3NS, and 3DS conditions were included in the model, and the effects of length and delay and their interaction were examined. In all cases, the reference group of the length variable was the shortest length, and it was DS condition for the delay variable. The parentheses in the first column give information on how that group differs from the reference group. If the regression is not tested on a given location, it is marked in gray. The numbers show the regression coefficients with asterisks marking the significance level (*** p < 0.001, ** p < 0.01, * p <0.05). If there was no significant effect, it is marked with dash (-).*

F0 (Hz)	**NP1**				
	Vpre	P	Vpost	R	F
2NS, 2DS, 3NS, 3DS					
length:delay			-		-
length (3)			-		0.98*
delay (NS)			3.29***		1.19**
1NS, 2NS, 3NS					
length (2NS)			5.79***		-1.34**
length (3NS)			5.65***		-
1NS, 2DS, 3DS					
length (2DS)			2.51***		-2.48***
length (3DS)			2.13***		-1.73***
1Pi, 2Pi, 3Pi (cf. 1NS/2DS/3DS, 2NS, 3NS)					
length (2Pi)	1.87**	3.94***		2.11**	
length (3Pi)	2.07**	5.06***		2.64**	

F0 (Hz)	**NP2**				
	Vpre	P	Vpost	R	F
2NS, 2DS, 3NS, 3DS					
length:delay	-	2.22*	-	2.17*	2.16*
length (3)	-	2.99	1.57***	2.33	1.46
delay (NS)	1.3***	0.53	0.67*	-0.77	-0.24

F0 (Hz)	NP3				
	Vpre	P	Vpost	R	F
3NS, 3DS					
delay	0.93*	1.3*	0.92*	-	-

F0 diff (Hz)	NP1-NP2				
	Vpre	P	Vpost	R	F
2NS, 2DS, 3NS, 3DS					
length:delay	-	-			-
length (3)	-	-3.46***			-1.83**
delay (NS)	-	2.78***			-

F0 diff (Hz)	NP2-NP3				
	Vpre	P	Vpost	R	F
3NS, 3DS					
delay	-	1.49*			

Table A 2. Results of the mixed-effects linear regressions on phrase and word durations. The rest of the information is identical to Table A 1.

Dur (ms)	**NP1**	**NP1-NP2**	**NP2**
	phrase dur	int dur	phrase dur
2NS, 2DS, 3NS, 3DS			
length:delay	-	-	-
length (3)	16.01***	12.01***	44.31***
delay (NS)	-	-6.86*	-
1NS, 2NS, 3NS			
length (2NS)	17.98***		
length (3NS)	40.03***		
1NS, 2DS, 3DS			
length (2DS)	20.24***		
length (3DS)	27.67***		

Dur (ms)	**NP1**			**NP1-NP2**	**NP2**		
	num1	col1	ani1	AND1	num2	col2	ani2
2NS, 2DS, 3NS, 3DS							
length:delay	-	-	-	-	-	-	-
length (3)	-	-	14.31***	8.53***	12.53***	-	29.45***
delay (NS)	-	-	-4.47*	-6.72***	-	-	-
1NS, 2NS, 3NS							
length (2NS)	-	-	15.41***				
length (3NS)	-	-	32.39***				
1NS, 2DS, 3DS							
length (2DS)	-	-	22.86***				
length (3DS)	-	-	31.82***				

REFERENCES

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, *3*, 255–309.

Boyce, S., & Menn, L. (1979). Peak vary, Endpoints don't: Implications for intonation theory. *Annual Meeting of the Berkeley Linguistics Society, 5*, 373–384.

Browman, C. P., & Goldstein, L. (1989). Articulatory Gestures as Phonological Units. *Phonology*, *6*(2), 201–251.

Browman, C. P., & Goldstein, L. (1990a). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 341–376). Cambridge: Cambridge University Press.

Browman, C. P., & Goldstein, L. (1990b). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, *18*(3), 299–320.

Bruce, G. (1982). Developing the Swedish intonation model. *Working papers/Lund University, Department of Linguistics and Phonetics*, *222*, 51–116.

Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America*, *103*(6), 3153–3161.

Byrd, D., & Krivokapić, J. (2021). Cracking Prosody in Articulatory Phonology. *Annual Review of Linguistics*, *7*, 31–53.

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

Chen, S. H., Liu, H., Xu, Y., & Larson, C. R. (2007). Voice F0 responses to pitch-shifted voice feedback during English speech. *The Journal of the Acoustical Society of America*, *121*(2), 1157–1163.

Christaller, J. G. (1875). *A grammar of the Asante and Fante language*. Basel: Basel Evangelical Missionary Society.

Clements, G. N. (1979). The Description of Terraced-Level Tone Languages. *Language*, *55*(3), 536-558.

Clements, G. N. (1990). The status of register in intonation theory: Comments on the papers by Ladd and by Inkelas and Leben. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 58–71). Cambridge: Cambridge University Press.

Cohen, A., & 't Hart, J. (1967). On the anatomy of intonation. *Lingua*, *19*, 177–192.

Collier, R. (1975). Physiological correlates of intonation patterns. *The Journal of the Acoustical Society of America*, *58*(1), 249–255.

Connell, B. (2001). Downdrift, Downstep, and Declination. *Proceedings of Typology of African Prosodic Systems Workshop*, *Bielefeld University, Germany*.

Connell, B. (2003). Pitch realization and the four tones of Mambila. In S. Kaji (Ed.), *Cross-linguistics studies of tonal phenomena* (pp. 181–197). Tokyo: Research Institute for the Languages and Cultures of Asia and Africa.

Connell, B. (2004). Tone, Utterance length and F0 scaling. *Proceedings of the International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*.

Connell, B. (2011). Downstep. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology*. John Wiley & Sons.

Connell, B., & Ladd, D. R. (1990). Aspects of pitch realisation in Yoruba. *Phonology*, *7*(1), 1–29.

Cooper, W. E., & Sorensen, J. M. (1981). *Fundamental frequency in sentence production*. New York, NY: Springer-Verlag.

Donath, T. M., Natke, U., & Kalveram, K. Th. (2002). Effects of frequency-shifted auditory feedback on voice F0 contours in syllables. *The Journal of the Acoustical Society of America*, *111*(1), 357–366.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(3), 545–572.

Fuchs, S., Petrone, C., Krivokapić, J., & Hoole, P. (2013). Acoustic and respiratory evidence for utterance planning in German. *Journal of Phonetics*, *41*(1), 29–47.

Fujisaki, H. (1983). Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. In P. F. MacNeilage (Ed.), *The Production of Speech* (pp. 39–55). New York, NY: Springer.

Fujisaki, H. (2003). Prosody, information, and modeling - with emphasis on tonal features of speech. *Workshop on Spoken Language Processing*.

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

Gao, M. (2008). *Mandarin Tones: An Articulatory Phonology Account*. PhD dissertation, Yale University.

Gårding, E. (1983). A Generative Model of Intonation. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and Measurements* (Vol. 14, pp. 11–25). Springer Berlin Heidelberg.

Hirschberg, J., & Pierrehumbert, J. (1986). The intonational structuring of discourse. In the *24th Annual Meeting of the Association for Computational Linguistics,* 136–144.

Inkelas, S., & Leben, W. R. (1990). Where phonology and phonetics intersect: The case of Hausa intonation. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 17–34). Cambridge: Cambridge University Press.

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *The Journal of the Acoustical Society of America*, *108*(3), 1246-1251.

Jones, J. A., & Munhall, K. G. (2002). The role of auditory feedback during phonation: Studies of Mandarin tone production. *Journal of Phonetics*, *30*(3), 303–320.

Katsika, A., Krivokapić, J., Mooshammer, C., Tiede, M., & Goldstein, L. (2014). The coordination of boundary tones and its interaction with prominence. *Journal of Phonetics*, *44*, 62–82.

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

Kochanski, G., Shih, C., & Jing, H. (2003). Quantitative measurement of prosodic strength in Mandarin. *Speech Communication*, *41*(4), 625–645.

Krivokapić, J. (2020). Prosody in Articulatory Phonology. In S. Shattuck-Hufnagel & J. Barnes (Eds.), *Prosodic Theory and Practice*. Cambridge, MA: MIT Press.

Krivokapić, J., Styler, W., & Parrell, B. (2020). Pause postures: The relationship between articulation and cognitive processes during pauses. *Journal of Phonetics*, *79*, 100953.

Ladd, D. R. (1983). Phonological Features of Intonational Peaks. *Language*, *59*(4), 721-759.

Ladd, D. R. (1984). Declination: A review and some hypotheses. *Phonology Yearbook*, *1*, 53–74.

Ladd, D. R. (1988). Declination ‘“reset”’ and the hierarchical organization of utterances. *The Journal of the Acoustical Society of America*, *84*(2), 530–544.

Ladd, D. R. (1990). Metrical representation of pitch register. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 35–57). Cambridge: Cambridge University Press.

Ladd, D. R. (1992). An introduction to intonational phonology. In G. J. Docherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (pp. 321–334). Cambridge: Cambridge University Press.

Ladd, D. R. (2008). *Intonational Phonology* (Second Edition). Cambridge: Cambridge University Press.

Ladd, D. R., & Johnson, C. (1987). “Metrical” factors in the scaling of sentence-initial accent peaks. *Phonetica*, *44*(4), 238–245.

Ladd, D. R., & Terken, J. (1995). Modeling intra- and inter-speaker pitch range variations. *Proceedings of the 13th International Congress of Phonetic Sciences*, 386–389.

Laniran, Y. O., & Clements, G. N. (2003). Downstep and high raising: Interacting factors in Yoruba tone production. *Journal of Phonetics*, *31*(2), 203–250.

Liberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. Oehrle (Eds.), *Language Sound Structure* (pp. 157–233). Cambridge, MA: MIT Press.

Lieberman, P. (1966). *Intonation, Perception, and Language*. PhD dissertation, Massachusetts Institute of Technology.

Maeda, S. (1976). *A Characterization of American English Intonation*. PhD dissertation, Massachusetts Institute of Technology.

Mücke, D., Nam, H., Hermes, A., & Goldstein, L. (2012). Coupling of tone and constriction gestures in pitch accents. In P. Hoole, L. Bombien, M. Pouplier, C. Mooshammer, & B. Kühnert (Eds.), *Consonant clusters and structural complexity* (pp. 205–229). Berlin: Walter de Gruyter.

Natke, U., Donath, T. M., & Kalveram, K. Th. (2003). Control of voice fundamental frequency in speaking versus singing. *The Journal of the Acoustical Society of America*, *113*(3), 1587–1593.

Natke, U., & Kalveram, K. T. (2001). Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables. *Journal of Speech, Language, and Hearing Research*, *44*, 577–584.

Niemann, H., Mücke, D., Nam, H., Goldstein, L., & Grice, M. (2011). Tones as gestures: The case of Italian and German. *IcPhS 2011*, 1486-1489.

Ohala, J. (1978). Production of tone. In V. A. Fromkin (Ed.), *Tone: A linguistic survey*. New York, NY: Academic Press.

Patel, R., Niziolek, C., Reilly, K., & Guenther, F. H. (2011). Prosodic adaptations to pitch perturbation in running speech. *Journal of Speech, Language, and Hearing Research*, *54*(4), 1051–1059.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. PhD dissertation, Massachusetts Institute of Technology.

Pierrehumbert, J. (1981). Synthesizing intonation. *The Journal of the Acoustical Society of America*, *70*(4), 985–995.

Pierrehumbert, J., & Beckman, M. E. (1988). *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Pike, K. L. (1945). *The intonation of American English*. Ann Arbor: University of Michigan Press.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Prieto, P., D'Imperio, M., Elordieta, G., Frota, S., & Vigário, M. (2006). Evidence for "soft" preplanning in tonal production: Initial scaling in Romance. *Speech Prosody 2006,* 803–806.

Prieto, P., Shih, C., & Nibert, H. (1996). Pitch downtrend in Spanish. *Journal of Phonetics*, *24*(4), 445–473.

Prom-on, S., Xu, Y., & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *The Journal of the Acoustical Society of America*, *125*(1), 405–424.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*(4), 333–382.

Saltzman, E., Nam, H., Krivokapić, J., & Goldstein, L. (2008). A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. *Proceedings of the 4th International Conference on Speech Prosody*, 175–184.

Scholz, F., & Chen, Y. (2014). Sentence planning and f0 scaling in Wenzhou Chinese. *Journal of Phonetics*, *47*, 81–91.

Shih, C. (2000). A Declination Model of Mandarin Chinese. In A. Botinis (Ed.), *Intonation* (Vol. 15, pp. 243–268). Springer Netherlands.

Shriberg, E., Ladd, D. R., Terken, J., & Stolcke, A. (1996). Modeling pitch range variation within and across speakers: Predicting F0 targets when "speaking up." *Proceedings of the 4th International Conference on Spoken Language Processing*, 1–4.

Sternberg, S., Knoll, R. L., Monsell, S., & Wright, C. E. (1988). Motor programs and hierarchical organization in the control of rapid speech. *Phonetica*, *34*, 175–197.

Thorsen, N. (1980). Intonation contours and stress group patterns in declarative sentences of varying length in ASC Danish. *Annual Report of the Institute of*

Phonetics University of Copenhagen, *14*, 1–29.

Thorsen, N. G. (1980). A study of the perception of sentence intonation—Evidence from Danish. *The Journal of the Acoustical Society of America*, *67*(3), 1014–1030.

Tilsen, S. (2007). Vowel-to-vowel coarticulation and dissimilation in response-priming. *UC Berkeley Phonology Lab Annual Report*, 416–458.

Tilsen, S. (2009). *Hierarchical spatiotemporal dynamics of speech rhythm and articulation*. PhD dissertation, University of California, Berkeley.

Tilsen, S. (2014). Selection and coordination of articulatory gestures in temporally constrained production. *Journal of Phonetics*, *44*, 26–46.

Tilsen, S. (2018). Three mechanisms for modeling articulation: Selection, coordination, and intention. *Cornell Working Papers in Phonetics and Phonology 2018*.

van den Berg, R., Gussenhoven, C., & Rietveld, T. (1992). Downstep in Dutch: Implications for a model. In G. J. Docherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (pp. 335–359). Cambridge: Cambridge University Press.

van Heuven, V. J. (2004). Planning in speech melody: Production and perception of downstep in Dutch. *LOT Occasional Series*, *2*, 83–93.

Ward, I. C. (1933). *The phonetic and tonal structure of Efik*. Cambridge: Heffer.

Whalen, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics*, *18*(1), 3–35.

Winston, F. D. (1960). The “mid” tone in Efik. *African Language Studies*, *1*, 185–192.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, *46*(3–4), 220–251.

Xu, Y., Larson, C. R., Bauer, J. J., & Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *The Journal of the Acoustical Society of America*, *116*(2), 1168–1178.

Xu, Y., & Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, *33*(4), 319–337.

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, *33*(2), 159–197.

Yi, H. (2017). *Lexical tone gestures*. PhD dissertation, Cornell University.

Yuan, J. (2004). *Intonation in Mandarin Chinese: Acoustics, Perception, and Computational Modeling*. PhD dissertation, Cornell University.