Online Control of Articulation Based on Auditory Feedback in Normal Speech and Stuttering: Behavioral and Modeling Studies

by

Shanqing Cai

B. Eng., Biomedical Engineering,
Tsinghua University, Beijing, China, 2005

M. S. E., Biomedical Engineering,
The Johns Hopkins University, Baltimore, Maryland, USA, 2007

Submitted to the Harvard-MIT Division of Health Science and Technology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY
IN SPEECH AND HEARING BIOSCIENCE AND TECHNOLOGY
AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February, 2012

Signature of Author _____
Harvard-MIT Division of Health Science and Technology
February, 2012

Certified by _____
Frank H. Guenther, Ph.D.
Professor of Speech, Language and Hearing Sciences and Biomedical Engineering, Boston University
Thesis Supervisor

Certified by _____
Joseph S. Perkell, Ph.D., D.M.D.
Senior Research Scientist, Research Laboratory of Electronics, MIT
Thesis Co-supervisor and Committee Chair

Accepted by _____
Ram Sasisekharan, Ph.D.
Edwin Hood Taplin Professor of Medical and Electrical Engineering
Director, Harvard-MIT Division of Health Science and Technology

Online Control of Articulation Based on Auditory Feedback in Normal Speech and Stuttering: Behavioral and Modeling Studies

by

Shanqing Cai

Submitted to the Harvard-MIT Division of Health Science and Technology
on February 3rd, 2012 in partial fulfillment of the requirements of the degree of
Doctor of Philosophy in Speech and Hearing Science and Technology

Abstract

Articulation of multisyllabic speech requires a high degree of accuracy in controlling the spatial (positional) and the temporal parameters of articulatory movements. In stuttering, a disorder of speech fluency, failures to meet these control requirements occur frequently, leading to dysfluencies such as sound repetitions and prolongations. Currently, little is known about the sensorimotor mechanisms underlying the control of multisyllabic articulation and how they break down in stuttering. This dissertation is focused on the interaction between multisyllabic articulation and auditory feedback (AF), the perception of one's own speech sounds during speech production, which has been shown previously to play important roles in quasi-static articulations as well as in the mechanisms of stuttering.

To investigate this topic empirically, we developed a digital signal processing platform for introducing flexible online perturbations of time-varying formants in speakers' AF during speech production. This platform was in a series of perturbation experiments, in which we aimed separately at elucidating the role of AF in controlling the spatial and temporal parameters of multisyllabic articulation. Under these perturbations of AF, normal subjects showed small but significant and specific online adjustments in the spatial and temporal parameters of articulation, which provided first evidence for a role of AF in the online fine-tuning of articulatory trajectories. To model and explain these findings, we designed and tested sqDIVA, a computational model for the sensory feedback-based control of speech movement timing. Test results indicated that this new model accurately accounted for the spatiotemporal compensation patterns observed in the perturbation experiments.

In addition, we investigated empirically how the AF-based online speech motor control differed between people who stutter (PWS) and normal speakers. The PWS group showed compensatory responses significantly smaller in magnitude and slower in onset compared to the control subjects' responses. This under-compensation to AF perturbation was observed for both quasi-static vowels and multisyllabic speech, and for both the spatial and temporal control of articulation. This abnormal sensorimotor performance supports the hypothesis that stuttering involves deficits in the rapid internal transformations between the auditory and motor domains, with important implications for the neural basis of this disorder.

Thesis Supervisor: Frank H. Guenther, Ph.D., Professor of Speech, Language and Hearing Sciences and Biomedical Engineering, Boston University

Thesis Co-supervisor: Joseph S. Perkell, Ph.D., D.M.D., Senior Research Scientist, Research Laboratory of Electronics, MIT

Thesis Committee Chair and Co-supervisor
Joseph S. Perkell, Ph.D., D.M.D.
Senior Research Scientist, Research Laboratory of Electronics, MIT

Thesis Supervisor:
Frank H. Guenther, Ph.D.
Professor of Speech, Language and Hearing Sciences and Biomedical Engineering, Boston University

Thesis Reader:
Michale S. Fee, Ph.D.
Professor of Neuroscience, Department of Brain and Cognitive Sciences, MIT

Acknowledgements

First of all, I would like to thank Dr. Joseph Perkell, who provided a tremendous amount of support to me throughout the four and half years of my PhD study. It was really fortunately for me to have an advisor like Joe who is willing to work with me as a colleague and cares about me. Dr. Frank Guenther, with his quick but deep and sharp thinking, guided me through my PhD study. He has been and will continue to be a role model for me in my academic career. Dr. Michale Fee, the reader of this dissertation, helped me to make this dissertation better by raising many important and thought-provoking questions.

Most of my time during this PhD study was spent at the MIT Speech Communication Group and the Boston University Speech Lab. It was in these two exceptional labs that I had the great opportunity to work with some really intelligent and supportive colleagues. Dr. Satrajit Ghosh helped me in virtually all aspects of my research, including the auditory-perturbation behavioral studies and the MRI data acquisition and analysis. I am most grateful to him for his generous sparing of his time to discuss with me about numerous data analysis and interpretation questions. Dr. Deryk Beal provided a great amount of help to my research, especially in the stuttering section of this dissertation. It is fair to say that without his expertise and help, that section would not be possible. Many members of the two labs gave me a lot of their time in helping me with MRI scans. These include Carrie Niziolek, Elisa Golfinopoulos, Jenn Segawa, Deryk Beal, Simon Overduin, Misha Panko, and Jason Tourville.

I really enjoyed sharing ideas and research tools with a lot of the researchers at the two labs, especially the journal club presentations and the discussions on various research topics. Oren Civier, another PhD student who worked on stuttering research at the BU Speech Lab, had many in-depth conversations with me about stuttering and modeling studies. He provided many helpful comments to the drafts of this thesis. Other people who shared their ideas and thoughts with me include Jason Tourville, Simon Overduin, Miriam Makhlouf, Jay Bohland, and Maya Peeva. Due to the space limit here, I cannot list everyone who stimulated my thinking, but I owe many of the ideas, hypothesis, and approaches in this thesis to discussions with them. I will miss the academically nurturing environment at these two labs on both sides of the Charles River.

Several affiliated members of the two labs also helped my intellectual growth. Dr. Dan Bullock and Dr. Stefanie Shattuck-Hufnagel both served on my oral qualification exam committee and discuss with me about my research projects, for which I am very grateful. Both of them engaged in in-depth discussions with me about various parts of this dissertation.

Dr. Nelson Kiang was supportive to me from the very beginning of my journey at MIT and SHBT. The advice and words of wisdom from him will continue to influence my research career

I thank the excellent technical and administrative support from the RLE staff, especially Seth Hall and Arlene Wint, who also made the MIT Speech Communication Group feel more like home.

I would also like to acknowledge the assistance of Ms. Adrianna DiGrande and Ms. Dianne Parris in the recruitment of the stuttering subjects.

This research was supported by a number of research funds, including NIH grants R01-DC0001925 and R56-DC0010849 (PI: J. Perkell) and NSF Doctoral Dissertation Grant #1056511. During my four and half years of PhD study, I have been supported by the MIT Edward Austin and Chyn Duog Shiah Memorial Graduate Fellowships, a Harvard Martinos Center Multimodal Neuroimaging Training Grant, and an ASA Raymond H. Stetson Fellowship in Speech Production and Phonetics.

Finally, I give my deepest thanks to my loving and lovely family: my wife Wei, my eight-month-old son David, my parents and parent-in-laws (especially mom-in-law Lu Xin and my mom Peihua Qiu, who came from China to the US to give Wei and me help raising David while both of us were in the final stages of our PhD studies). Without the tremendous support and sacrifice on their part, none of this would have been possible.

Table of Contents

List of Figures

(Continued)	
4.1	Simulation of the relations between the feedback weight α_{FB} of DIVA and the model's compensatory response to perturbation of the AF of F1
4.2	Rationale for preserving the data from the two PWS subjects who had mild hearing losses according to our hearing-screening criterion
4.3	An example of the labeling of the onset and offset of the nucleus vowel [ɛ] in the word "pet"
4.4	Compensatory responses under the randomize F1 perturbations
4.5	Differences in the F1 trajectories produced under the perturbation (Down and Up) conditions with short or long spacing after the preceding perturbation trial
4.6	The compensatory F1 changes under the subset-mode analysis
4.7	Average composite response curves from the PWS (red) and PFS (black) groups
4.8	Comparison of online compensation between the PFS and PWS groups
4.9	Latency of the compensatory responses
4.10	Auditory acuity to changes in F1 of the vowel [ɛ] and its relation to the magnitude of the compensation to perturbation in the PWS and PFS groups
4.11	F2 compensation curves under the Up (A) and Down (B) perturbations
4.12	Changes in time intervals during the utterance "I owe you a yo-yo" under the Up (red) and Down (blue) perturbations from the noPert baseline
4.13	F2 trajectories produced by the PFS and PWS subjects and their changes from the no-perturbation baseline under the time-varying temporal (Accel and Decel) perturbations
4.14	Comparison of the online timing corrections under the temporal (Accel and Decel) perturbations in the PFS and PWS groups
4.15	Changes in time intervals during the utterance "I owe you a yo-yo" under the Accel (orange) and Decel (Green) perturbations from the noPert baseline
4.16	Comparison of the variability of F1 production between the two groups
4.17	Comparing the latency of response fitted with the sqDIVA-T model to the timing-perturbation responses from the control (left) and PWS (right) subjects.
4.18	A schematic diagram illustrating our hypothesis regarding the mechanisms underlying the transitions between different syllables in a multisyllabic speech utterance

List of Tables

Abbreviations

In alphabetical order

2AFC	Two-alternative-force-choice
Accel	Accelerating perturbation (A subtype of temporal perturbation. See Sect. 2.3)
AF	Auditory feedback
ANOVA	Analysis of variance
AVS	Average vowel spacing
BA#	Brodmann's area # (# being a positive integer, e.g., 44)
BG	Basal ganglia
BOLD	Blood oxygen level dependent
CBF	Cerebral blood flow
CI	Cochlear implant
CRC	Composite response curve
CPG	Central pattern generator
CQ	Competitive queuing
CWS	Children who stutter
CV	Consonant-vowel
CVC	Consonant-vowel-consonant
DAF	Delayed auditory feedback
dBA	Decibel, A-weighted
Decel	Decelerating perturbation (One subtype of temporal perturbation. See Sect. 2.3)
DIVA	Directions to Velocities of the Articulators (A neurocomputational model of speech motor control)
DSP	Digital signal processing or processor
DOF	Degree of freedom
DPS	Duration pattern sequence
DTI	Diffusion tensor imaging
EEG	Electroencephalography
EFR	Early following response (see Sect. 4.2.2)
F0	Fundamental frequency (pitch) of voice
F1	First formant frequency
F2	Second formant frequency
FDR	False Discovery Rate
fMRI	Functional MRI
GODIVA	Gradient-order DIVA (A neurocomputational model of syllable planning and sequencing in speech production)
HL	Hearing level
HSD	Honestly Significant Differences (Tukey's test)
IFG	Inferior frontal gyrus

IM	Internal model
JND	Just noticeable difference
LSE	Least square error
MEG	Magnetoencephalography
MIS	Motor induced suppression (of auditory responses)
noPert	No-perturbation
PDS	Persistent developmental stuttering
PFS	Person(s) with fluent speech
PPI	Psychophysiological interaction
pSTG	Posterior superior temporal gyrus
PT	Planum temporale
PWS	Person(s) who stutter(s)
rCBF	Regional cerebral blood flow
RLE	Research Laboratory of Electronics
RM-ANOVA	Repeated measures analysis of variance
RMS	Root-mean-square
SEM	1. Standard error of mean 2. Structural equation modeling (Should be clear from context)
SF	Somatosensory feedback
SFC	State feedback control (a model)
SMA	Supplementary motor area
SPL	Sound pressure level
Spt	Sylvian fissure at the parietal-temporal junction
SSM	Speech sound map
STI	Spatiotemporal index
sqDIVA	Sequential DIVA (a new model)
TD	Task Dynamic (a model of speech motor control)
TMS	Transcranial magnetic stimulation
VOT	Voice onset time
vMC	Ventral motor cortex
vPMC	Ventral premotor cortex
VBM	Voxel-based morphometry
WM	White matter

Chapter 1. Introduction

Speech is one of the most important means of interpersonal communication and arguably one of the most complicated motor skills mastered by most humans. The production of speech is a complex process involving multiple stages of formulation, transformation and control (Levelt 1989). This dissertation focuses on the last stage in this process: articulation, namely the transformation of phonetic plan of an utterance through movements of articulators (e.g., jaw, tongue and lips) into sequences of speech sounds that can be heard and understood by a listener.

To produce speech, the brain mobilizes a large number of organs and muscles of the upper body (Barlow and Andreatta 1999). These structures can be divided roughly into three groups based on their positions relative to the vocal folds (glottis). The subglottal structures, including the muscles of the chest wall and the diaphragm, provide the airflow support for speech. Aero-dynamically induced vibration of the vocal folds, which are stiffened under the control of laryngeal muscles, generates the source sound for voiced sounds such as vowels[1]. Supraglottal structures, of which the jaw, the tongue, the velum and the lips are the main components, modulate the shape of the vocal tract and thereby change the acoustic transfer function of the vocal tract from the glottis to the mouth opening (Stevens 1998). Hence the spectral envelopes of speech sounds, which determine the identity of the phonemes in languages such as English, are controlled primarily by the supraglottal structures. This thesis will focus on the supraglottal mechanisms of articulation.

During speech production, the brain is also faced with the challenge of control precision. For certain sounds, a deviation in the position of an articulator by a few millimeters can completely alter the spectrum of the sound and result in mispronunciation of the intended sound (Stevens 1998). This is what we refer to as the *spatial* aspect of the speech motor control.

[1] The sound sources for other types of speech sounds, such as alveolar fricatives, are generated not by the vocal folds, but by supraglottal structures. But this dissertation will focus primarily on vowel sounds.

The normal speaking rate of American English is approximately 9 - 14 phonemes per second (Crystal and House 1988). The transitions between phonemes and syllables need to be precisely timed and smooth in order for the produced speech to be fluent and intelligible. In addition, speech prosody, which conveys syntactic and paralinguistic information such as attitude and emotional state of the speaker, employs subtle changes in segmental and suprasegmental timing. The timing of the phonemes and syllables and the transitions between them is what we refer to as *temporal* parameters of speech motor control.

Considering the multiplicative relation between the spatial and temporal complexity of speech, the articulation of fluent speech is a remarkable motor ability. Yet for people with normal speaking ability, articulating fluent speech usually feels automatic and effortless. How does the brain achieve such rapid and precise control of scores of muscles to produce intelligible speech? How does the brain acquire the ability to do this? These are the central question pursued by researchers in the field of speech motor control and its neurophysiology.

However, articulation can break down in various disorders of speech production. Persistent developmental stuttering[2] is a developmental disorder of speech fluency affecting approximately 1% of the adult population and 5% of children (Bloodstein and Ratner 2008; Chang 2011). This disorder is characterized by frequent disruptions of fluent speech by involuntary sound and syllable repetitions, prolongations and blocks. The etiology of stuttering remains unclear but is an active area of research. Which parts of the speech motor system are abnormal in people who stutter? What kinds of functional breakdown lead to dysfluencies in this disorder? These are the questions that not only are pursued by researchers of stuttering in hope of obtaining a better understanding of the etiology of this disorder (which will eventually lead to improved diagnosis and treatment of it), but also draw keen interest from the area of normal speech physiology, since as in many other fields of physiology, how the speech motor system breaks down may shed important lights on how the system functions normally.

[2] For the sake of simplicity, we will refer to this disorder as "stuttering" below. This ignores the distinction between persistent developmental stuttering and other types of stuttering, such as neurogenic stuttering (e.g., Helm 1978). But this should not be problematic here because we are not concerned with the latter in this dissertation.

1.1. The role of auditory feedback in speech motor control

In the past two decades, speech scientists have directed much attention toward the role of auditory feedback in speech motor control. Auditory feedback (AF) refers to the speech signals heard by the speaker himself or herself when speaking. For a number of reasons, AF has been an interesting and valuable research topic in understanding speech motor control. First, AF is the most accessible and easy-to-manipulate part of the speech chain. The speech faculty is unique to humans and there is a lack of widely accepted animal models for investigating the control of speech production. This renders many powerful neurophysiological tools, such as invasive *in vivo* recording of neuronal activities inaccessible to speech neurophysiologists. But thanks to audio and digital signal processing technologies, researchers can easily measure and manipulate AF during speech. Second, manipulation of AF permits establishment of causal relations among different events and components of the speech process. Neuroimaging tools (e.g., functional MRI and PET) and motion measurement tools such as electromagnetic articulography (EMA, Perkell et al. 1992) are limited by their correlational nature, despite the invaluable insights into the speech motor system afforded by these correlational methods (Indefrey et al. 2004). As observational methods, by themselves they cannot be used to infer the causal relations among different processes in the speech system. In contrast, by manipulating AF in experimentally controlled ways and measuring the effects of such manipulations, causal relations between auditory perception and speech motor control can be established. Third, evidence and theoretical arguments have been accumulating in support of the view that the goal of speech articulation may largely reside in AF. We will review such empirical and theoretical developments in the following sections.

In the past several decades, a remarkable number of empirical observations have been made on the role of AF in speech motor control. Thanks to these data, theoreticians have reached a number of shared conclusions regarding how AF is utilized by the speech motor system to

achieve more intelligible speech. However, considerable controversy remains in certain key issues about the role of AF in speech motor control. The key issues of debate are

1) whether the AF plays a role in the online, moment-by-moment control of speech;
2) if so, what the specific role played by AF is and how important this role is;
3) whether speech disorders such as stuttering are due to, or at least accompanied by, abnormalities in this putative AF-based online control of speech.

These are the issues that will be addressed by the experiments in this thesis. In the following sections, I will develop a critical review of what we currently know about the role of AF in speech motor control, what we do not know, and what hinders our progress toward a deeper and more thorough understanding of this topic. From this review, I will derive the motivation and justification for the main aims of this thesis.

1.1.1. The role of auditory feedback revealed by hearing loss and cochlear implant usage

It has long been known that functional hearing is required for the acquisition of speech motor skills. Children with moderate-to-profound prelingual hearing impairment usually fail to develop intelligible speech (Hudgins and Numbers 1942; Gold 1980). The deficits in deaf children's speech encompass almost all aspects of voicing, articulation and prosody (Boone 1966; Calvert 1961; Nober 1967; Markides 1970; Stark and Levitt 1974; Parkhurst and Levitt 1978). The restoration of functional hearing with cochlear implants (CIs) can lead to considerable benefit for the acquisition of speech production in prelingually deaf children (e.g., Osberger et al. 1993; Tye-Murray and Kirk 1993; Tye-Murray and Spencer 1995).

These observations clearly show that hearing is closely related to the "root" of the speech motor system, i.e., it is a critical part of the sensorimotor apparatus required for acquiring speech motor skills. However, it should be noted that a prelingual loss of hearing deprives the child of both the AF of self-produced speech and the perception of speech uttered by others, i.e., the

models to learn from. Therefore it is generally unclear to what degree the speech motor deficits in prelingual deafness should be attributed to the unavailability of AF and to what degree they should be attributed to the unavailability of models for learning (speech produced by adults in the environment).

It is difficult to obtain direct empirical observations on the role of AF in the acquisition of speech skills since there are no known disorders that affect AF alone while sparing the audition of external speech sounds. However, information regarding the roles of AF in the *maintenance* of speech motor skills and in the online, moment-to-moment *control* of speech articulation is available from several lines of investigation. These include the effects of postlingual hearing loss, the effects of masking noise and other manipulation of AF, and the more recent investigations based on perturbations of AF using more advanced digital audio signal processing techniques.

Although postlingual hearing loss generally does not lead to significant loss of the ability to produce intelligible speech, previous studies have revealed deterioration of various aspects of speech motor control in people suffering from postlingual hearing loss. As a coarse but ecologically relevant measure of speech motor performance, the intelligibility of speech has been found to be compromised by post-lingual hearing loss (Cowie et al. 1982) and to benefit from subsequent restoration of functional hearing by CIs (Gould et al. 2001).

Systematic investigations of the details of speech motor skill deterioration following post-lingual deafening have also been carried out. First, previous findings indicate that hearing plays an important role in the maintenance of quality of *vowel* production. The contrasts between the formant frequencies of different vowels (e.g., measured as the size of the vowel spaces, or average vowel separation, AVS) have been shown to decrease following post-lingual hearing loss (Waldstein 1990; Perkell et al. 2001; Ménard et al. 2007; Lane et al. 2007) and increase back toward normative values under the functional hearing afforded by CIs (Ménard et al. 2007; Lane et al. 2007). Second, previous studies have shown similar effects of hearing loss and CIs on the quality of consonant production, such as the spectral contrasts between alveolar fricatives /s/ and

/ʃ/, laterals (/r/) and the voice onset time (VOT) of stop consonants (Waldstein 1990; Lane et al. 1994; Matthies et al. 1996; Lane et al. 2007).

From the findings reviewed above, we can infer that although AF is not absolutely required for producing fluent speech in adults who have already acquired speech motor skills, it may play important roles in maintaining the overall intelligibility of speech and the quality of many important aspects of segmental and suprasegmental speech articulation. However, these studies can provide few, if any, clues to the role of AF in the online, moment-to-moment control of articulation, the primary focus of the current dissertation. This is due to the following reasons. First, the long-term longitudinal approach focuses on the gradual, plastic changes in speech motor skills and speaks little to the online control mechanisms. On the other hand, the "within-session on-off" approach used by those studies employed changes in hearing status on a relatively long time scale since the cochlear implants were switched on or off between relatively long blocks of trials. Therefore it is possible that 1) adaptation and adjustment to speech may occur quickly over the course of a small number of trials (c.f., Figure 6 of Matthies et al. 1996), confounding the contribution the within-trial, online control mechanism. In addition, it is also possible that with experience, the CI users have developed two separate sets of motor internal models (IMs) for speaking without AF (as before implantation or when CI is off) and with partially restored AF (as when CI is on). For such subjects, hearing status may constitute a "motor context", and a simple change in the hearing status during the experiment, which the subjects were most likely aware of, or even the mere anticipation of a change in hearing status, may be sufficient to activate certain changes in speech motor strategies, allowing switching from the use of one set of IMs to another (Wolpert and Kawato 1998).

One of the ways in which the above limitations can be bypassed in order to investigate the role of AF in the online control of speech motor control is to switch the CI suddenly and unexpectedly on or off during the course of an utterance. This approach was taken by Perkell et al. (2007), who measured the changes in vowel and sibilant contrasts, vowel duration, intensity and F0 following sudden switching on and off of the CI while the CI user produced nonsensical

[$dV_1n\ C_1V_2d$] utterances such as "don-shed". Vowel and sibilant contrast distances were not found to be significantly affected by sudden changes in hearing status, but vowel durations showed systematic and significant changes when AF (through CI) was switched suddenly off. Specifically, sudden switching off the CI led to significant increases in the vowel durations. This held true for both the first vowel (V_1), during which the switching occurred, and the following one (V_2).

This intriguing finding by Perkell et al. (2007) regarding the segment duration changes due to sudden changes in hearing status provides some support for the role of AF in the online control of utterance timing, i.e., the temporal parameters of multisyllabic articulation[3]. However, does the lack of vowel and sibilant contrast in that study indicate that AF is not involved in the online control of the articulation positions (i.e., the spatial parameters) of those speech sounds? This would be an erroneous conclusion to draw. Despite being sudden and unanticipated, the intra-utterance turning off and on of AF used by Perkell et al. remains a coarse manipulation of auditory feedback, which may be too gross in granularity to be a precise probe of the involvement of AF in the online control of the spatial parameters of articulation. To form a possibly not-so-appropriate analogy, imagine a group of aliens, upon their first arrival on Earth, trying to find out how an Earth car works. Their observation that suddenly pulling the steering wheel out from the steering shaft when a car is running fails to alter the direction of the car's travel can hardly be used as evidence that the steering wheel is unrelated to controlling the car's direction. A more subtle and relevant manipulation of the system, such as turning the steering wheel, would be far more revealing as to the function of the steering wheel. Perturbations of AF using digital signal processing techniques, which are used in the current study, constitute this kind of more elegant and informative manipulation. Previous studies based on similar methods will be reviewed Section 1.1.4.

[3] This statement is based on the assumption that the findings under such an unnatural condition as a sudden loss of AF due to the off-switching of a CI can be extrapolated to speech under normal conditions, the validity of which will be explored in Section 1.1.3.

1.1.2. The role of auditory feedback revealed by masking and delayed auditory feedback

Apart from switching AF on and off in hearing-impaired CI users, there are other more routinely accessible methods of manipulating AF. These include the masking of AF using intense noise (called "noise masking") and delaying the air-conducted AF by a certain amount of time, typically between 20 and 500 ms (called delayed auditory feedback, or DAF). These methods can be applied easily to normal-hearing subjects. Noise masking and DAF are similar to the CI methods in being coarse and lacking in specificity with respect to acoustic parameters and spatiotemporal granularity. Masking the AF of normal-hearing speakers with noise has been shown to have systematic effects on various aspects of speech articulation, including increasing voice intensity, decreasing speaking rate, and altering several segmental features such as the phoneme contrasts (Lane et al. 1970; Lane and Tranel 1971; Van Summers et al. 1988; Perkell et al. 2007).

Another technique for manipulating AF that become available relatively early is delayed auditory feedback, namely delaying AF by a fraction of a second before playing it back. Also, this type of manipulation has large effects on normal speaker's speech. These observations are perhaps why delayed auditory feedback (DAF) was the first manipulation of AF that drew widespread attention from researchers. The original discoverer of the phenomenon, Bernard S. Lee (1950, 1951) described the subjects' oral reading under delays of 40 – 280 ms as either 1) slower than normal and or 2) if not slower than normal, containing halts and repetitions of syllables and words. Because of latter, Lee referred to this phenomenon as "artificial stutter". The disfluencies and speech errors that occur under DAF have also been carefully documented. These include prolongation of the syllable-medial vowels (Fairbanks 1955), repetition of syllables and words (Atkinson 1953; Fairbanks and Guttman 1958), segmental errors and substitutions (Atkinson 1953), and sound omissions (Korowbow 1955). Zimmerman et al. (1988) reported interesting timing relations between AF and articulatory movements in fluent and disfluent syllables produced by normal speakers under DAF. Specifically, they found that during

fluent speech under these DAF conditions, the offset of the AF from the syllable that preceded a fluently produced syllable occurred predominantly *before* the onset of the fluently produced syllable. By contrast, when audible disfluencies (e.g., repetitions) or inaudible articulatory breakdowns occurred, this above inter-syllabic timing relation between AF and articulatory movement was often violated. These findings seem to indicate that under DAF, the speech motor system adopts a strategy of waiting for the AF from the previous syllable to finish before initiating the following one in order to avoid potential conflicts and subsequent disfluencies[4].

More detailed and systematic studies of DAF followed the seminal papers by Lee. It was found that the speaking rate showed a nonmonotonic relation with the amount of delay. Specifically, it decreases with increasing amount of delay for short delays, but subsequently increases with further increased delays. The minimum speaking rate occurs at approximately 180 ms of delay (Black 1951; Atkinson 1953; Fairbanks 1955; Sussman 1971; Zimmerman 1988). Interestingly, Peters (1954) and Davidson (1959) observed that decreasing the delay of AF from its normal physiological value, about 1 ms, to 0.15-0.3 ms using an electronic devicef led to small (~2-4%) but significant increases in speaking rate in oral reading tasks. This was essentially the opposite phenomenon to DAF.

Unfortunately, to our knowledge, no prior studies have been carried out on the effects of DAF on acoustic contrasts between sounds (e.g., AVS and /s/-/ʃ/ spectral contrast) and on articulatory or acoustic variability. Knowledge about these effects could contribute to our understanding of these phoneme and the more generally of the role of AF in speech motor control. In the current dissertation, we will endeavor to gain knowledge in this regard by using a technique different than DAF, namely more granular and well-controlled manipulation of AF, which will be described in detail in Chapter 2.

[4] Note that in normal speech, the auditory feedback from the proceeding vowel always ends before the release of the following stop consonant, a pattern which is consistent with the timing relation seen most frequently during the fluent speech under DAF.

1.1.3. Investigations on auditory feedback based on perturbation techniques

The above-reviewed methods for investigating the relation between hearing and speech production suffer from several limitations when their results are used to infer the role of AF in speech motor control. First, as mentioned earlier, the methods of studying speech in the hearing impaired cannot easily separate the consequences of the loss of AF and those of the loss of hearing others' speech. As for noise masking and delayed auditory feedback, there is a considerable amount of debate concerning whether the results obtained under such grossly altered AF states can be extrapolated to speech production under ordinary, unperturbed conditions (Lane and Tranel 1971; Borden 1979). Thirdly, these methods affect all parts of speech equally and it is difficult to direct them toward specific temporal or spectral windows (but see Perkell et al. 2007).

Fortunately, a more elegant type of manipulations of AF, which can at least partly overcome these shortcomings of older methods, has become possible thanks to advances in digital signal processing technology. By using such techniques, individual parameters of speech AF, such as formant frequencies, pitch (F0), intensity and even the timing of individual syllables and phonemes can be manipulated in specific and well-controlled ways, without grossly altering the natural pattern of AF or causing conscious awareness on the part of the subject.

Before the 1980s, the only acoustic parameters that could be manipulated in near-real time[5] were intensity and delay. This is why most of the studies on AF and speech prior to 1980 either used DAF (already discussed in Sect. 1.1.2) or manipulated the intensity of the vocal feedback. Siegel and Pick (1974) observed that when AF was amplified by 10 or 20 dB relative to its natural intensity, subjects compensated for the perturbation by lowering the level of their produced speech by a small but significant amount, about 10% of the AF change. This effect was basically the opposite of the Lombard effect (Lane and Tranel 1971).

[5] In the context of this thesis, we use the term "real time" to refer to processing delays shorter than 15 – 20 ms, that is, delays in AF that are unnoticeable to the subject and for all practical purposes do not elicit DAF effects.

Generally speaking, two categories of experiment designs are used to study how perturbations of AF affect speech production. The first category of designs, which we dub *sustained perturbation paradigms*, involves aggregating trials with perturbations in one continuous part (i.e., block) of the experiment and trials without perturbations in other blocks. In this paradigm, the subject is repeatedly exposed to the same perturbation in the perturbed block. As a consequence, the subject usually develops an *adaptation* to the perturbed AF environment showing long-term changes in his or her production. These changes can be verified by measuring the aftereffect, namely the difference between the subject's production immediately before and after the cessation of perturbation. If the subject's production shows an aftereffect, i.e., the adaptive response does not immediately disappear after the perturbation has been removed, it constitutes the most convincing evidence for motor adaptation.

The second type of design, which is more relevant to the purpose of the current thesis, is called the *randomized perturbation paradigm*. Unlike the sustained perturbation design, this type of paradigm focuses on the role of AF in the online, moment-to-moment control of ongoing speech. The basic strategy is to compare the speech produced by the subject in the perturbed and the unperturbed (i.e., baseline) trials. However, motor adaptation, which could possibly occur if the subject is presented with too many consecutive perturbed trials (as in the sustained perturbation paradigm) or if the perturbation comes at predictable times (e.g., at regular intervals), becomes a confounding factor. In order to minimize the likelihood of motor adaptation, this paradigm randomly intersperses the perturbed and baseline trials and uses substantially fewer perturbed trials than baseline trials. Certain important parameters of online control, such as the latency of the feedback loop, can only be investigated by using the randomized perturbation paradigm.

Researchers have used both methods to study AF-based control of speech production. However, since the randomized paradigm was used by most previous F0 studies and will be used by this current thesis, we will focus on studies that used the randomized paradigm and discuss the sustained perturbation paradigm only within the context of formant feedback control.

There is consensus among different studies conducted in the past three decades that despite some trial-to-trial and inter-subject variability, the prevalent pattern of response to a small and often unnoticeable perturbation to the AF of F0 is a counteracting change in the produced F0 of the subject (Elman 1981; Burnett et al. 1998; Larson et al. 2000; Burnett and Larson 2002; Natke and Kalveram 2001; Donath et al. 2002; Natke et al. 2003; Bauer and Larson 2003; Jones and Munhall 2002; Xu et al. 2004; Chen et al. 2007; Larson et al. 2008). This response, which is often referred to as the compensatory response, has been shown to have a latency in the range of 75 to 240 ms[6]. The magnitude of the response, also varies from study to study, but generally falls into the range of 10 – 55% of the AF pitch shift[7]. There are many possible reasons for this large study-to-study variation in response latency and in the ratio of compensation. Among these are differences in speech task and experimental procedure. A number of studies have shown that both the latency and magnitude of the compensatory response can be systematically modulated by task-related and procedural factors including duration (Burnett et al. 1998) and onset abruptness (Larson et al. 2000) of the pitch-shift stimulus, as well as whether the F0 target is static or time-varying (Burnett and Larson 2002; Chen et al. 2007; Xu et al. 2004)

Apart from the studies by Xu, Chen and colleagues, a few other studies have also examined the role of AF in the control of F0 in multisyllabic utterances. Natke and Kalvarem (2001) introduced perturbation to the AF of pitch while German speakers produced two utterances with different stress patterns. In one of the nonsensical utterances ['ta:tatas], the first syllable was stressed and long in duration; in the other utterance [ta'ta:tas], the second syllable, instead of the first one, was stressed. The onset of the F0 shift always occurred during the first syllable and the shift always lasted for the duration of the entire utterance. It was shown that when the first

[6] In addition to difference in task and experiment design, the wide range of latencies reported in those studies is also due to the different methods for calculating the latency of compensatory response. Some studies used ±2 SD from the mean as the threshold for determining latency (e.g., Jones and Munhall 2002); others used p-value thresholds from t-tests (e.g., Xu et al. 2004); still others used the Siegel-Castellan change point test (e.g., Donath et al. 2002). The first two methods are more conservative than the third one and are sensitive to changes in sample size and measurement noise, and hence tend to generate overestimated response latencies.

[7] 100% corresponds to a full compensation: i.e., a change in F0 production which brings the F0 in the AF back to its pre-perturbation value.

syllable was stressed and long, there was a significant same-syllable compensatory adjustment in the produced F0. However, when the first syllable was unstressed and short, within-syllable compensation did not occur. These findings show that the latency of the compensatory response, mostly likely determined by the synaptic and other information processing delays in the central nervous system, is an important constraint in the online feedback-based control of speech production.

Natke and Kalvarem (2001) also showed that no matter whether the first syllable showed a compensatory F0 adjustment or not, the second syllable always showed F0 changes that opposed the perturbation in the first syllable. Similarly, Donath et al. (2002) used ['ta:tatas] as the stimulus utterance and demonstrated the interesting observation that compensatory F0 changes could be observed even in an unperturbed trial that is preceded immediately by a perturbed trial. These persistent compensation across syllables and trials may be similar to aftereffects in speech motor adaptation and will be important considerations in designing and analyzing studies based on the randomized perturbation paradigm.

Perturbation studies on formants of vowels

The technique for real-time manipulation of the formant frequencies in speech sounds did not emerge until 1997, when John Houde created a DSP-based hardware that could shift the first and second formant frequencies of vowels (Houde and Jordan 1998; 2002). Influenced by the tradition of research on limb motor adaptation at their institution, MIT (e.g., Shadmehr and Mussa-Ivaldi 1994), Houde and Jordan (1998; 2002) used a sustained perturbation paradigm to investigate how speakers adapt to repeated perturbations to AF of the F1 and F2 of the vowel [ɛ]. Their main findings can be summarized as follows.

1) On average, subjects compensated for 54% of the AF perturbation, although considerable variability in the compensatory response was observed across subjects. While some

subjects compensated fully, some others barely showed any changes in production under the perturbation.

2) After the compensatory response had developed, it persisted even if AF is temporarily blocked by masking noise. This indicates that the observed changes in the subjects' production were not due primarily to online corrections, but mainly due to genuine alterations in the motor program for the vowel, viz., adaptation.

To avoid possible interference from bone-conducted AF, Houde and Jordan instructed subjects to produce whispered speech. Findings similar to the above were also made by other groups of researchers on voiced speech (Purcell and Munhall 2006a; Villacorta et al. 2007; Munhall et al. 2009; MacDonald et al. 2010). However, the ratio of compensation found by those later researchers were smaller than that found by Houde and Jordan (2002) (Villacorta et al. 2007: 40%; Purcell and Munhall 2006b: 11-16%; MacDonald et al. 2010: 15-20%). The reason for this discrepancy in compensation magnitude remains unclear, but may be attributable to factors such as the absolute magnitude of the formant perturbation (MacDonald et al. 2010), the conformity of the formant shifts to the subject-specific vowel space (Niziolek et al. 2010), and the effect of bone-conducted AF.

The adaptive changes in speech motor programs in response to perturbations of AF have recently been extended from vowel formants to other acoustic parameters and other types of phonemes, including the alveolar fricative [ʃ] and (Shiller et al. 2009) and the VOTs of the stop consonants [t] and [d] (Mitsuya et al. 2010).

Unlike research on F0 perturbation (see the preceding section), only a small number of studies to date (Purcell and Munhall 2006b; Tourville et al. 2008) have used the randomized paradigm to examine the role of AF in the *online* control of formants or other acoustic parameters associated with supraglottal articulators. In Purcell and Munhall's (2006b) experiment, subjects sustained the isolated English vowels [ɪ], [ɛ] and [æ] for more than 1000 ms. In randomly selected 5% of the trials, AF of the F1 of the vowels were shifted up or down

unbeknownst to the subjects. In the perturbed trials, the subjects on average showed within-trial compensatory change in their produced F1 values in the direction opposite to that of the AF perturbation. The magnitude of this compensatory response increased monotonically with increasing delay from the onset of the F1 shift. The ratio was 2-3% at 275 ms following shift onset, but reached as high as 10 -15% at 1000 ms after shift onset. This response was qualitatively similar to the pitch compensation seen in F0 perturbation studies (reviewed in the preceding section). However, this study could not provide direct information about the latency of this response because a ramped perturbation onset, rather than a sudden (square-wave) onset was used.

Purcell and Munhall (2006b) used a quite unnatural speech task: to sustain a vowel in isolation (i.e., without syllabic or lexical context) for a long duration. Therefore it was not entirely clear how generalizable their findings are to more usual type of speech at normal speaking rates. Behavioral data from the fMRI study by Tourville and colleagues (2008) partially addressed this issue. Tourville et al. (2008) instructed subjects to produce monosyllabic CVC words that contain the monophthongs [ɛ] while inside an MRI scanner. The subjects were instructed to utter those words in a slightly prolonged way (not as prolonged as in Purcell and Munhall 2006b). A compensatory F1 adjustment in the subjects' production similar to that observed in Purcell and Munhall (2006b) was found, indicating that online AF-based control of vowel articulation does occur during both sustained vocalization and word production. The latency of the response was calculated to be approximately 170 ms, within the range of pitch-shift compensation latencies found in previous studies (see the preceding subsection). The magnitude of the compensatory response found by Tourville and colleagues was about 5-6% of that of the AF perturbation. This ratio is smaller than the maximum ratio of compensation reported by Purcell and Munhall (2006b), possibly due to the shorter vowel duration used by Tourville et al. (2008). Considering that the vowel duration in Tourville et al. (2008) is closer to the duration of single vocalic phonemes in real-life speech than that in Purcell and Munhall (2006b), we may conclude that the role of AF in the within-phoneme and within-syllable control

of articulation certainly exists but is quite small in magnitude, perhaps even smaller than the role of AF in controlling F0 (e.g., Natke and Kalveram 2001; Donath et al. 2002; Natke et al. 2003; Xu et al. 2004; Chen et al. 2007).

Both Purcell and Munhall (2006b) and Tourville et al. (2008) used monosyllabic utterances: either isolated vowels or CVC words. To our knowledge, no previous studies have been conducted of how the speech motor system uses AF to control formants (or other non-F0 acoustic parameters) during time-varying speech sounds or multisyllabic speech[8]. Therefore knowledge regarding the role of AF in the online control of multisyllabic articulation is still lacking. The current dissertation aims to fill this gap.

1.1.4. The role of somatosensory feedback in speech motor control and its relation to auditory feedback

Apart from auditory feedback, somatosensory feedback (SF) is another major channel of sensory feedback available to the brain during speech production. The roles of SF in both the online and offline feedback control of speech articulation have been demonstrated by previous studies.

The technique for online mechanical perturbation of the vocal tract during speech articulation was developed in the late 1970s, much earlier than the emergence of the digital signal processing techniques for near real-time fine-grained manipulation of auditory feedback. A series of pioneer investigations by James Abbs, Vincent Gracco and their colleagues shed light on how alterations in somatosensory feedback information affect the execution of speech movements.

The details of the methodology differ among studies (Folkins and Abbs 1975; Abbs and Gracco 1983; Abbs and Gracco 1984; Gracco and Abbs 1985; Gracco and Abbs 1989), but they share a common pattern as follows. The subjects produce a short utterance consisting of a

[8] Cai et al. (2010) investigated the AF-based sensorimotor adaptation in the production the triphthong /iau/ in Mandarin. But as a study based on the sustained perturbation paradigm, it did not shed light upon the online AF-based control of the production of time-varying speech sounds.

bilabial stop consonant (e.g., [aba], "sappaple" and "hapap"). Shortly before the onset of the bilabial closure movement for the [b] or [p] sound, an electromechanical device is activated (usually automatically) to exert a downward mechanical force on the jaw or the lower lip and hence to cause disturbance to the bilabial closure. This kind of perturbations were introduced in a randomly selected subset of trials; the remaining trials were produced in the absence of perturbation. The percentage of perturbed trials was kept low (~15%) in order to reduce the likelihood of anticipation or adaptation. Under this type of mechanical perturbation, subjects' articulations were observed to exhibit short-latency adjustments that ensured successful contact of the lips. These compensatory adjustments were manifested as increased movement magnitudes and electromyographic (EMG) signals of both the lower lip (the perturbed articulator per se) and the upper lip (the unperturbed but task-related articulator). Specifically, the downward movement (depression) of the upper lip increases significantly, despite the fact that the upper lip received no external perturbing forces. The compensatory changes in movement magnitude were shown to be comparable to the passive displacements caused by the mechanical perturbations, hence the "ratio of compensation" can be said to be much greater under these mechanical perturbations than under auditory perturbation. In addition to changes in the muscle activities and the amplitude of the articulatory movements, timing of the syllables following the perturbation could also be altered as a consequence of the perturbation and/or the compensation to the perturbation (Gracco and Abbs 1989). The latency of these compensatory responses fall into the range of 25 – 80 ms, which argues against the brainstem-mediated perioral reflex, and is consistent with a suprabulbar, long-loop pathway that involves the motor cortex and the cerebellum (Gracco and Abbs 1985).

However, these findings did not shed light upon the nature of the articulatory goals: whether they are specified in the articulatory/somatosensory domain (c.f., the tract variables in the Task Dynamics model, Saltzman and Munhall 1989) or the acoustic/auditory domain (c.f., Perkell et al. 1997; Guenther et al. 1998; Perkell 2011), because the compensatory articulatory adjustments helped to attain both the normal vocal tract constriction (e.g., contact of the lips) and the normal

acoustic outcome (e.g., the stop-associated sudden-onset followed by the plosive burst) in the face of the mechanical perturbation. A number of studies based on the sustained mechanical perturbation of the jaw have provided some support for goals in the somatosensory domain (Tremblay et al. 2003; Nasir and Ostry 2008; Tremblay et al. 2008) by showing that subjects made nearly complete (100%) adaptive changes in response to mechanical force perturbations that have no significant effects on the acoustic parameters of speech.

However, other studies based on sustained mechanical or geometrical perturbation of the vocal tract have provided different insights to this issue. Savariaux et al. (1995) showed that when a tube is inserted between the lips to prevent the normal rounding of the lips during the French rounded vowel [u], seven out of 11 subjects altered the position of the tongue to a more posterior position in order to completely or partially preserve the F1/F2 pattern of the vowel, indicating that the speech motor system is capable of ignoring the usual articulatory positions, when necessary, in order to achieve a desired acoustic/auditory outcome of articulation. In an elegant and informative recent study, Feng and colleagues (2011) introduced simultaneous perturbations to AF and SF when subjects produced the vowels [ɛ] and [æ] embedded in CVC words. The AF perturbation involved an upward shift of F1; the SF perturbation involved an upward force applied by a robotic device to the jaw. Note that this combination of cross-modality perturbation was incongruent, in the sense that there was no way in which a single articulatory change can minimize errors in both the auditory and somatosensory domains. For example, an upward movement of the jaw and the tongue can reduce the auditory error, but it is at the expense of increased deviation in the SF domain. Conversely, a downward movement of the jaw and the tongue reduces SF but would result in further increased error in the AF domain. Feng and colleagues showed that on average, the subjects adopted the former strategy, i.e., they elevated the jaw and tongue in order to reduce the AF error, at the expense of increased SF error.

In summary, these findings seem to indicate that speech motor control involves a hierarchy of goals, in which the auditory goals are at the higher level and the somatosensory goals are on a

lower one. When errors in the AF and SF domains are in conflict, the brain will choose to attend to the task of minimizing the AF error with a higher priority. However, when the compensation for the SF error does not involve any increased AF error, somatosensory compensation will be implemented. These conclusions are largely consistent with the DIVA model of speech motor control, which we will review in Section 1.2.2.

Although most of the experiments cited above are concerned with single syllables or words, it is important to move to investigations of multi-syllabic speech stimuli in order to ensure the relevance of experimental results to typical speaking situations. More attention has been devoted to multisyllabic speech by investigations into somatosensory-motor interactions than by investigations on auditory-motor interaction, perhaps because of less difficult technical challenges involved. Some evidence has been gathered for a definite role of SF in the online control of spatiotemporal parameters of multisyllabic speech (e.g., Gracco and Abbs 1989). However, such a role of SF may not be directly extrapolated directly to AF without empirical confirmation, because of the considerable differences in the anatomy and neurophysiology of the auditory and somatosensory systems as well as their functional roles in the speech motor system (Perkell 1997; Guenther et al. 1998; Guenther 2006). In addition, none of the previous studies of SF-based online control of articulation used linguistically meaningful or multi-word utterances, so that the scope of the findings in this area and their generalizability to natural speech remain limited. In this dissertation study, we aim to overcome these limitations by using a new AF perturbation paradigm.

1.2. Models of the sensorimotor processes underlying speech articulation

1.2.1. The Task Dynamic model

The Task Dynamic (TD) model (Saltzman and Munhall 1989), developed at the Haskins Laboratories, is a highly influential model of speech motor control. This model is closely related to the speech motor control framework known as Articulatory Phonology (Browman and

Goldstein 1992). This model is based on the implicit premise that the goals of articulation are to reach targets in the domain of vocal tract constrictions (or targets of *tract variables*, in the language of the 1989 paper). Each phoneme in a language is hypothesized to be associated with a specific target for tract variable(s) (e.g., zero lip aperture for the bilabial stop [b], anterior and superior position of tongue dorsum constriction for the vowel [i], etc.) In a hierarchical way, the targets for tract variables direct the movements of the *model articulators*, which are correlates of the physical articulators, such as the jaw, the tongue tip, the tongue body, the lips, etc. The behavior of the model articulators is modeled with second-order ordinary differential equations, analogous to a mass-spring model in which the mass, stiffness and damping are the fundamental parameters. The tract variable target for each individual phoneme is a static set of values in the parameter space and constitutes a "point attractor" for the model articulator positions. As such, the kinematic trajectories seen in the simulations of the TD/AP model are an emergent property of the second-order linear mechanical system. This end-point-only control paradigm has its root in equilibrium-point theory of general motor control (cf., Feldman 1966, 1986; Perrier et al. 1996), and is at odds with theories of whole-trajectory planning (e.g., Flash and Hogan 1985).

Apart from the tract variable targets, which are timeless spatial configurations used for implementing individual phonemes, the TD model also includes a level of timing control, referred to as the *intergestural* level. This intergestural level is essentially a "score" for the temporal organization of the sequence of phonemes in an utterance. Each phoneme is associated with a gestural *activation variable*, a single scalar that is a function of time. This activation variable specifies the strength with which each tract variable target of a phoneme attempts to influence the shape of the vocal tract at any given point in time. In the original construction of TD, these activation variables took a discrete 0-1 value and were specified in an *ad hoc* manner (i.e., "hard-wired"). In later developments of the model (Nam and Saltzman 2003), the activation variables could take continuous values and become controlled by coupled oscillators. However, despite these attempts at modeling timing phenomena in speech in a more mechanistic and less

"hard-wired" manner, timing in the TD model continues to be largely feedforward and hence lacking in details about feedback interactions.

1.2.2. The DIVA model

The DIVA[9] model is another influential model of speech motor control. Unlike the Task Dynamics/Articulatory Phonology model, DIVA is based on the hypothesis that speech movements are planned and controlled primarily in the auditory perceptual domain. In their 1998 publication on DIVA (Guenther et al. 1998), Guenther and colleagues provided eloquent arguments for the auditory-reference-frame hypothesis and argued against the central tenet of the Task Dynamic model, i.e., that speech movements are planned within the reference frame of the constrictions of the vocal tract. Their argument was developed from several angles. First, as a theoretical consideration, it was pointed out that during speech development, the brain doesn't have a reliable source of teaching signals for learning the mapping between motor commands and vocal tract constrictions[10]. However, this teaching is indispensible for the formation of a reliable internal model, which is necessary for generating movement commands based on the current state of the speech motor system. By contrast, teaching signals for a mapping between auditory signal and motor command are readily available; they are the difference between auditory percepts of the self-produced speech sound and auditory percepts of the speech sounds produced by others (e.g., speech sounds produced by parents or caretakers, the model for learning). Second, various empirical findings argue against speech motor planning in a constriction frame. These include the articulatory trading relations between lip rounding and tongue body raising for the vowel [u] (Perkell et al. 1993), the compensatory tongue position changes in subjects attempting to produce the vowel [u] with a tube inserted between their lips (preventing sufficient rounding of the lips) (Savariaux et al. 1995), and the relative acoustic

[9] DIVA is the shorthand for "Directions into the Velocities of the Articulators".
[10] Note that proprioception can only provide the brain with information about current state of the vocal tract constrictions, but it cannot provide information about whether this constraint is appropriate for the production of the sound being produced.

invariance of the American English [r] produced by wide ranges of tongue shapes/positions and vocal tract configurations (Guenther et al. 1999). It was further demonstrated that the DIVA model, which carries out its movement planning in the auditory frame, was capable of simulating all the above findings.

These arguments formed the basis for later developments of the DIVA model, which is entirely on an auditory frame of speech motor planning. The most up-to-date revision of the DIVA model has been described in Guenther et al. (2006) and Golfinopoulos et al. (2009). Figure 1.1 is a schematic of this model.

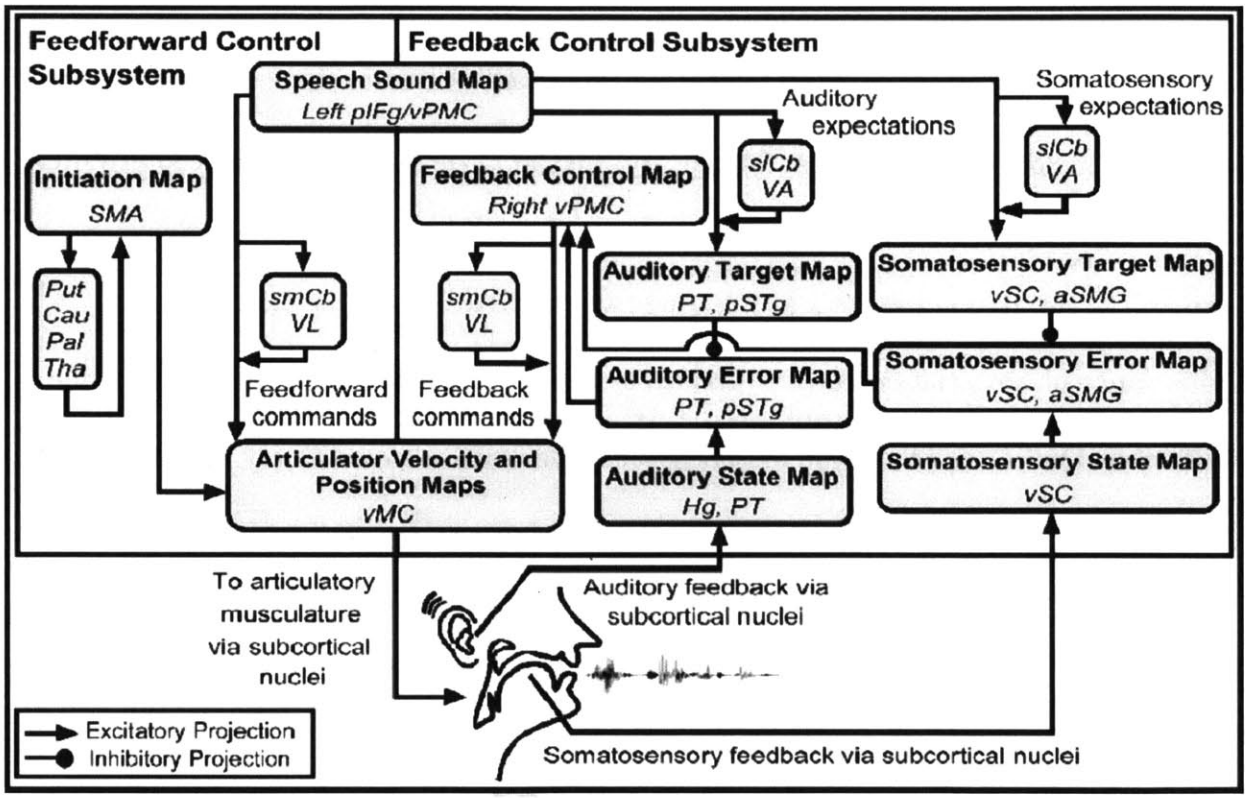


Figure. 1.1. A schematic diagram of the most up-to-date version of the DIVA model. (Reproduced from Golfinopoulus et al. 2009).

As can be seen in Fig. 1.1, the DIVA model spans many levels of the neural and neuromuscular system, from the prefrontal cerebral cortical areas to the muscle groups of the articulators. On the lowest level, the DIVA model incorporates a vocal-tract model slightly modified from the Maeda synthesizer (Maeda 1982). This vocal tract model has 8 degrees of

freedom, including dimensions of articulator geometry such as jaw height, tongue height and tongue shape. The movement (i.e., changes of spatial position in time) in each articulatory dimension is governed by the activities of a pair of antagonist motor command neurons. This leads to a 16-dimension motor command vector. This is the form of the motor commands issued from the primary motor cortex (ventral Motor Cortex, vMC) to the model vocal tract.

In order to compute these motor commands, the primary motor cortex integrates the function of two pathways which operate in parallel: the feedforward pathway and the feedback pathway. This feedforward-feedback integration is described by the following equation[11]:

$$M(t) = M(0) + \alpha_{FF} \int_0^t \dot{M}_{FF}(t) g(t) dt + \alpha_{FB} \int_0^t \dot{M}_{FB}(t) g(t) dt , \quad (1.1)$$

$M(t)$ is the motor state (position) at time t. $\dot{M}_{FF}$ and $\dot{M}_{FB}$ are respectively the feedforward and feedback motor commands, in the form of a velocity (i.e., rate of change) of the articulatory position. The two parameters, α_{FF} and α_{FB} are the relative weights of the feedback and feedforward pathways and they are constrained to sum to 1. The greater the value of α_{FB}, the smaller the value of α_{FF}, the greater responses the model will rely on auditory feedback for speech motor control.

The separation of the feedforward and feedback pathways is a key feature of the DIVA model. In its "mature" state, e.g., in an adult speaker, the model assigns a much greater value to the feedforward pathway (α_{FF} = 0.85) than to the feedback pathway (α_{FF} = 0.15) (Tourville et al. 2008). This makes the model capable of producing speech sounds with the feedforward pathway alone, e.g., when AF and SF are both blocked by masking and anesthesia (Ringel and Steer 1963). This feedforward-biased setup is also consistent with findings from auditory perturbation studies showing that the compensatory adjustments made to the articulations in

[11] This equation is based on Equation (2) of Guenther et al. (2006).

response to AF perturbations are generally incomplete and account for less than 20-30% of the perturbation (e.g., Purcell and Munhall 2006b; Tourville e al. 2008).

During production of a previously learned syllable[12], the Speech Sound Map (SSM, hypothesized to be located in the left ventral premotor cortex, vPMC) reads out a set of pre-learned feedforward motor commands for the syllable. These feedforward motor commands (i.e., Z_{PM}) are essentially a spatiotemporal "score" for articulatory movements during the production of this syllable. This score differs from the "point attractor" mode of control in the TD model in that it specifies the temporal details of how the articulatory movements should unfold. The feedforward motor command for a syllable is learned through feedback-based correction during the first few attempts in producing this syllable. The "teaching signal" for this learning comes from the auditory error signals, i.e., mismatches between the auditory target or goal of the sound and actual auditory feedback during the production. Hence it can be seen that auditory target regions are the primarily goal for motor planning in the DIVA model. The auditory target regions in DIVA are finite-width formant-value intervals that evolve in time, instead of being point targets (which would appear as lines, instead of bands in Fig. 1.2). This feature helps to account for many well-known empirical findings in speech production research, ranging from carryover coarticulation, contextual variability and rate effects (Guenther 1995).

[12] To be precise, the units of production in the DIVA model are frequently produced phoneme, syllables, and short words (Guenther et al. 2006). We will refer to the units as syllables for the sake of simplicity.

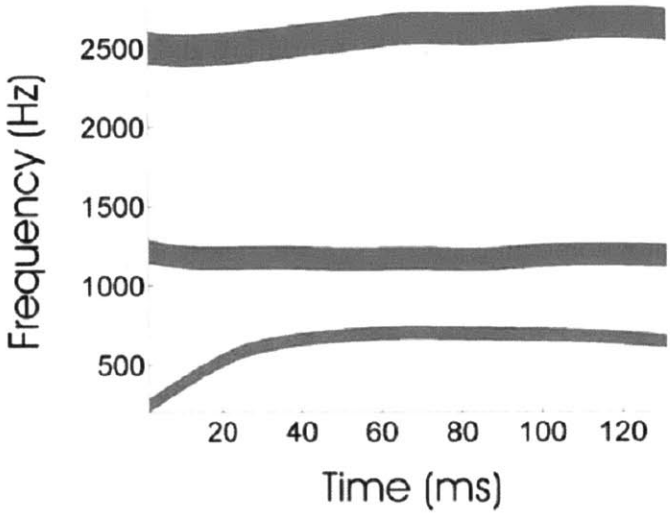


Figure 1.2. An example of the auditory goal region in DIVA for the syllable [ba]. The three gray bands show how the targets for F1, F2 and F3 (from bottom to top) evolve over time during the production of this syllable. (Reproduced from Guenther 2006)

During the first attempt at producing a syllable, the feedforward motor commands are set at an arbitrary configuration. This will obviously lead to misses of the auditory target. Hence, an error signals are generated. These error signals are generated as a consequence of the comparison that takes place in the auditory error map (Fig. 1.1), which continuously monitors the mismatches between auditory goal regions and the actual auditory feedback. The auditory goal regions are supplied by the SSM to the auditory error map via the auditory target map when the production of a syllable is initiated. The auditory error signals are then transformed to corrective motor commands by the inverse IMs or the feedback control map. These corrective motor commands serve two purposes. First, as a part of the motor learning process (i.e., practice), they will be incorporated into the feedforward motor command for the syllable being produced, so that future productions of the syllable will come closer to the auditory target. Second, they are incorporated into the ongoing motor commands with a short latency (see Equation 1.1), which serves the purpose of online feedback-based correction. Therefore we can see that corrective motor commands plays critical roles both in the "tuning up" of the feedforward pathway and in the proper functioning of the feedback control pathway. As such, the computation of the

corrective motor commands is probably the most important component of DIVA. This computation is succinctly described by the following equation:

$$\dot{M}_{FB}(t) = \Delta Au(t - \tau_{AuM}) \cdot z_{AuM}, \qquad (1.2)^{13}$$

wherein ΔAu is the auditory error, τ_{AuM} is a neural transmission delay, and z_{AuM} is a transformation matrix learned through "babbling", i.e., early period of motor experimentation based on random movements. z_{AuM} can be thought of as an inverse internal model that translates vectors in the auditory (formant domain) into the motor domain. The z_{AuM} weights take a general, nonlinear form, which captures the nonlinear and complex relation between vocal tract configuration and formant frequencies. In the most recent development of DIVA, the right ventral premotor cortex (vPMC, the right homolog of SSM) is postulated to be the seat of the feedback control mechanisms (Golfinopoulus et al. 2009). This cortical area computes and sends corrective motor commands to the vMC (partly) via the superior medial cerebellum and the ventrolateral nucleus of the thalamus. This localization is based on the fMRI results of increased activation in the right vPMC during randomized perturbation of vowel formants (Tourville et al. 2008) and increased activation in a similar area during randomized mechanical perturbation to the vocal tract during production (Golfinopoulus et al. 2011).This delineates the division of labor between the left and right hemispheres in the DIVA model, in which SSM in the left hemisphere is the center of the feedforward pathway and the feedback control map in the right hemisphere is the "hub" of the feedback pathway.

The DIVA model requires a few (6 – 8) practice tokens to master a typical syllable, that is, to produce the syllable with perceptually negligible amounts of acoustic error (Guenther 2006). Once a syllable is well learned, the model automatically forms a somatosensory expectation for this syllable. During subsequent productions of the syllable, somatosensory feedback from the vocal tract will be compared with this learned expectation in the "somatosensory error map" of

[13] This equation is based on Equation (9) of Guenther et al. (2006). For the sake of simplicity, and because we are not concerned with somatosensory perturbations and responses in this dissertation, the somatosensory component of the corrective motor command is omitted from this equation.

the model (See Fig. 1.1). Mismatches between the feedback and the expectation lead to somatosensory errors. In a way similar to the auditory error-based control, these error signals will be used to generate corrective motor commands using a set of somatosensory-to-motor weights. This feature endows the DIVA model with the ability to accurately simulate findings from the somatosensory perturbation-adaptation studies (e.g., Tremblay et al. 2003, 2008; Nasir and Ostry 2008; Feng et al. 2011). However, unlike the Task Dynamic model, the DIVA model treats somatosensory expectations as secondary to auditory goals, which are regarded the primary goals of speech motor planning.

1.2.3. Comparison of the DIVA and Task Dynamic models

As two competing models of speech motor control, DIVA and TD differ in many key aspects. Perhaps the most important difference between the two models is the reference frame in which speech movements are planned. The TD model implicitly assumes a tract-variable (or constriction) frame of reference. This property of TD is inherited from the linguistic theory of phonological features (see Stevens 1998[14]), which are mostly defined in the articulatory domain (e.g., tongue heights for vowels). The DIVA model explicitly assumes an auditory frame of movement planning and defends this idea with a number of convincing arguments (see Sect. 1.2.2). Thus it is clear that these two models have different targets of articulation. The auditory targets of DIVA enables it to accurately simulate compensatory responses to various kinds of auditory perturbation, a feature the TD model does not have because of its lack of consideration of the auditory domain.

The second significant difference between the two models is the attention paid to learning. The TD model is primarily a performance model. That is, it is applicable mainly to the mature state of the speech motor system and is not relevant to the process by which the movements are

[14] However, it needs to be pointed out that Stevens was among the first people to stress the importance of the acoustic properties of speech sounds for understanding speech motor control and to systematically study the relations between speech acoustics and articulation.

learned. The tract variable targets in the model are "tuned" in an *ad hoc* way (Saltzman and Munhall 1989), and it is unclear what biologically plausible processes may underlie such tuning. By contrast, one of the most important features of the DIVA model is its ability to model the processes in which novel syllables are learned. In fact, DIVA's capacity to learn is closely related to its choice of auditory frame as the frame of planning, taking AF as the teaching signals.

These two models also are different in biological plausibility. To date, authors of the TD model have not published systematic attempts to localize different components of the TD model to specific brain regions. This model is primarily oriented toward explaining behavioral data. In contrast, since the 2006 publication (Guenther 2006), the DIVA group attempted to make the model as neuroanatomically and neurophysiologically plausible as possible. As Fig. 1.1 shows, all major components of this model are localized to specific cortical or subcortical regions. These localizations are based on neuroimaging results from the same group (e.g., Bohland et al. 2006; Tourville et al. 2008; Peeva et al. 2010; Golfinopoulus et al. 2011) and other groups working on neuroimaging studies of speech production. Despite there being many tentative and potentially controversial details in the localization within the model (e.g., the omission of brainstem nuclei related to speech motor functions, the role of SMA in motor initiation and timing, etc.), DIVA is to our knowledge the most biologically detailed and plausible model of speech motor control currently in existence.

From a motor behavioral perspective, DIVA and TD differ in another key aspect. Auditory targets in DIVA are specified as time-varying regions in the auditory (formant) domain. These targets specify on a moment-by-moment basis how the formants should evolve with time. As a consequence, the learned articulatory motor commands in DIVA also specify on a moment-by-moment basis how the velocity of the articulators should evolve with time. Hence we can say that DIVA is a "trajectory-planning" model. By contrast, TD does not plan the spatiotemporal details of the articulatory trajectories; it only plans the static point attractors (equilibrium states) of the articulators. The articulator trajectories in TD are not results of planning, but results of the interaction between these point attractors, the level of the activation variables (see Sect. 1.2.1)

and the dynamic properties of the model articulators. It is still unclear at this point which model is closer to the speech motor system in reality. But the Cai et al. (2010a, 2010b) findings concerning the auditory-motor adaptation in the formant trajectory of Mandarin triphghons (a type of time-varying vowels) indicate that the speech motor system does compensate for perturbations to the AF of formant trajectories for time-varying vowels, which support the trajectory planning paradigm in DIVA.

Despite these differences, both models share a number of shortcomings. First, their treatments of articulatory timing leave a lot to be desired. The DIVA model is primarily concerned with the articulation of single units, such as syllables and frequently used short words (i.e., units in the mental syllabary, Levelt and Wheeldon 1994), and devotes little attention to how the transitional articulation between these elements are controlled and how timing patterns emerge in multisyllabic utterances. The TD model devotes relatively more attention to these topics, but adopts only a hard-wired approach, instead of a mechanistic one, to this problem.

In reviewing the currently most influential computational models of speech motor control, another recent model merits a brief mention. The GODIVA model (Bohland et al. 2009) is a computationally explicit model of the process underlying the sequencing of the *neural representation* of multiple syllables in an utterance (e.g., [tɑ] and [pɪk] in the utterance "topic"). This model is based on the frame-content theory of speech planning (e.g., Shattuck-Hufnagel 1987; MacNeilage 1998) and hypothesizes that the pre-supplementary motor area (pre-SMA) contains a neural-map representation of the syllable frames (e.g., CV, CVC) and the inferior frontal sulcus contains a neural-map representation of the phonemes (e.g, [t], [ɑ], etc.). A central issue in sequential motor control is the mechanism that ensures the correct ordering of the constituent elements. In the GODIVA model, the order of the syllables and phonemes is represented with a primacy gradient (e.g., Rhodes et al. 2004), in which greater model neuron activity signifies earlier positions in a sequence. Through a two-layer neural network with lateral and recurrent inhibition called competitive queuing (CQ, Bullock and Rhodes 2003, Rhodes et al.

2004), the primary-gradient representation is converted into a sequence of activations in the syllabic-frame and phonemic representation that evolves in an orderly fashion in time. The syllabic-frame and phonemic representations are integrated through a module thought to correspond to the basal ganglia to generate an ordered activation of syllable units in the SSM of the GODIVA model, intended to be the interface with the DIVA model.

GODIVA is a powerful and insightful model for syllabic sequencing in speech and its hypothesis are well-grounded in neuroimaging findings. However, the GODIVA model lacks components for the articulatory and sensorimotor process of speech production. Therefore it is a model that resides purely on the cognitive level and doesn't speak to the sensorimotor level of speech production. As such, GODIVA and DIVA are complementary. The two models need to be integrated to achieve a quantitative and comprehensive model of the processes underlying multisyllabic speech articulation.

An important issue in the future integration of DIVA and GODIVA is the role played by sensory feedback (including AF and SF) in sequential motor execution during multisyllabic uttearnces. In Chapter 3 of this dissertation, we will make a first attempt in this direction by constructing a computational model of the interaction between AF and motor execution and timing during a multisyllabic speech utterance. This model will be rooted in the basic premises and formulations of the DIVA model but will be extended from single-syllable to multisyllabic articulation. The results from the psychophysical experiments in Chapter 2 will be used as constraints for the tuning of this model.

1.3. Stuttering and the possible implications of sensory feedback

1.3.1. Overview of Stuttering and Sensorimotor Functions in this Disorder

Stuttering is a disorder of speech fluency characterized by frequent interruption of the normal flow of speech by various types of dysfluencies, including repetitions of syllables or sounds, prolongations of sounds, silent blockages and broken words. Stuttering is a developmental

disorder with the age of onset between the age of 2 and 6 years, during early speech acquisition (Bloodstein and Ratner 2008). Childhood stuttering has a high rate of spontaneous recovery (60 – 80%), therefore the incidence in children (~5%) is higher than the prevalence in adults (~1%, Culton 1986; Porfert and Rosenfield; Mansson 2000). The cases of unrecovered childhood stuttering in adults are called persistent developmental stuttering (PDS). Since this dissertation is concerned with adults who stutter, we will use the two terms "PDS" and "stuttering" interchangeably. Males are 3 – 5 times more likely to have PDS than females (e.g., Kidd et al. 1978; Yairi and Ambrose 2005; Craig and Tran 2005).

Scientific investigation of stuttering has yet to reveal the etiology of this disorder. But it has been long known that the cause of stuttering includes a genetic component (e.g., Howie et al. 1981; Ambrose et al. 1997; Riaz et al. 2005; Suresh et al. 2006; Wittke-Thompson et al. 2007). Recent advances in genetic research have shown that stuttering (or at least certain subtypes of it) are related to several genes, ranging from genes involved in intracellular lysosomal function (Kang et al. 2010) to genes that encode dopaminergic receptors (Wang et al. 2009). However, the detailed biological pathway leading from the abnormal genotypes to the behavioral characteristics of stuttering is far from being understood.

In this thesis, we will explore the motor component of this disorder and specifically focus on the interaction between online sensory feedback and speech motor control in stuttering, which we believe to be at the core of the etiology of this disorder. The assumption is consistent with a sizeable corpus of evidence. This includes atypical performance by people who stutter (PWS) in many sensorimotor tasks. For example, stutterers are slower than controls in initiating speech and non-speech movements upon detecting a sensory (visual or auditory) "go" cue (e.g., Adams and Hayden 1976; Cross and Luper 1987; Ferrand et al. 1991; Jones et al. 2002). When performing non-speech motor tasks that require fast, online monitoring of time-varying sensory information, PWS show slower and/or less accurate performance compared to normal controls (Stark and Pierce 1970; Neilson and Neilson 1979; Nudelman et al. 1987). Loucks et al. (2007) observed that stuttering adults showed substantially worse accuracy on a jaw-phonatory

coordination task than fluent controls. Loucks and de Nil (2006) showed that PWS were less accurate and more variable in their performance than fluent controls on a jaw target reaching task, but this between-group difference decreased when stutterers recieved the aid of the visual feedback of their jaw positions.

Other observations point specifically to a direct involvement of AF in leading to moments of dysfluencies in stuttering. Noise masking of AF is one of the most well-known fluency-inducing conditions for PWS. The reduction of stuttering frequency under loud masking noise has been shown during oral reading and conversation by various studies (e.g., Cherry and Sayers 1956; Maraist and Hutton 1957; Sutton and Chase 1961; Conture and Brayton 1975; Hutchinson and Norris 1977; Martin et al. 1984, 1985; Stager et al. 1997)[15]. Anecdotally, there have been reports of recovery from stuttering following onsets of profound hearing loss in adulthood (e.g., Van Riper 1982, p. 383).

DAF, a manipulation of AF that causes breakdown of fluency in normal speakers (see Sect. 1.1.2), paradoxically leads to significant and instantaneous improvement of the fluency in many PWS (e.g., Webster et al. 1970; Hutchinson and Norris 1977; Stephen and Haggard 1980; Stager et al. 1997; Kalinowski et al. 1993)[16]. In addition, several studies have shown that some stutterers speak more fluently under manipulations of auditory feedback other than DAF. These include frequency shifted AF (e.g., Kalinowski et al. 1993, Ingham et al. 1997), in which the entire spectrum of the AF is shifted up or down by around 0.25 - 0.5 octaves.

[15] The percentage reduction in the frequency of dysfluencies under noise masking reported in previous studies covers a wide range, from 20% to 100% (complete elimination of stuttering events). It is task-dependent (oral reading or spontaneous speech) and varies from person to person.

[16] However, it should be noted that there have been also reports of DAF *increasing* the dysfluency of PWS (Hayden et al. 1977), an effect similar to the effect of DAF on normal speakers. As pointed out by Bloodstein and Ratner (2008, p. 299), a common report is that mild stutterers typically show responses to DAF that are similar to normal speakers, while moderate and severe stutterers show improvement in fluency under DAF.

1.3.2. Models and hypotheses about the relations between sensory feedback and stuttering

The above-reviewed findings on the effects of AF manipulation on the fluency of stutterers have led several authors to propose that the primary cause of stuttering is the existence of certain abnormal relations between the auditory system and the speech motor system.

Neilson and Neilson (1987) argued that stuttering is a result of defective internal models (IMs), specifically a faiulre to form stable and accurate inverse IMs that transform desired acoustic output into articulatory movements or to effectively use those inverse IMs during speech. According to this hypothesis, AF is not used to control speech movements in a closed-loop (i.e., servo control) way, but is instead used in checking and updating the inverse IMs responsible for the generating of articulatory trajectories. They interpreted the fluency enhancing effect of the noise masking as a consequence of reduced task demand (c.f. Starkweather 1987). In their model, AF is utilized by the brain to adaptively and continuously update the inverse IMs when mismatches between the desired auditory output and the sensory feedback arise[17]. This adaptive updating consumes neural computational "resources" and may exhaust the limited supply of "neural computational capacity" in stutterers and thereby lead to poor performance in other tasks, such as using the IMs to compute the articulatory trajectories for the speech sounds, which in turn leads to dysfluencies. When AF is masked by noise, the computational load of updating the IM is reduced or eliminated, liberating limited neural resources for better functioning of the inverse IMs, thus contributing to improved fluency.

Postma and Kolk (1993) proposed that stuttering results from "covert repairing" of articulatory plans, or in other words, the internal editing of prepared motor programs of articulation following predictions of errors that have not occurred yet. Their idea of the internal

[17] It should be noted that this model by Neilson and Neilson (1987) is similar to the DIVA model (Guenther et al. 1998; 2006; Golfinopoulus et al. 2010) in that its mature (adult) version it doesn't require AF to produce speech movements. But whereas the Neilson and Neilson model uses AF to adaptively update inverse IMs, the DIVA model uses AF for somwhat different purposes, including 1) updating of syllabic motor programs and 2) online correction of speech movements. The account of stuttering developed by Civier (2010) is also similar to the theory of Neilson and Neilson (1987) in many regards.

monitoring process in speech production is rooted in the model of Levelt (1989). They pointed out that since covert repairing is a process that takes time (like any other functions of the brain), when covert repairing takes place, it often leads to interruptions of speech flow. They went further to argue that PWS are less able than nonstutterers to generate error-free articulatory plans for certain reasons, and thus are faced with higher-than-normal frequency of covert repair and the interruptions of the flow of speech.

Mysak (1960) proposed the servo theory of stuttering. His theory was based on the servo theory of normal speech production (Fairbanks 1954), in which sensory feedback is compared with an "input signal" to generate error signals that contribute to driving the movement of the articulators. Mysak's (1959) model diagram was an extension of Fairbank's (1954). In Mysak's (1959) model, the motor commands driving the articulators come from two sources, a) the linguistic "inputs" and b) the corrective motor commands that originate from the mismatch between those inputs and the sensory feedback. In this regard, Mysak (1959) model remotely resembles the much later DIVA model, which contains two control pathways that contribute simultaneously to the final motor commands. Based on the 1959 model, Mysak (1960) argued that disruptions in any part of the system or the connections within it may result in "verbalizing deautomaticity" and breakdowns of speech fluency. In his model, disruptions of either the pathway that doesn't rely on sensory feedback or the pathway that does may result in (different subtypes of) stuttering. With regard to the role of AF, one of the types of disruptions mentioned by Mysak was an abnormality in the "sensory units". He proposed that the air- or bone-conduction pathways for AF of speech may be abnormal in some stutterers, leading to abnormally long feedback delay. In this regard, stuttering can be regarded as a form of "naturally occurring DAF effect". However, it is unclear how this theory may explain the observation that adding extra latency to the AF can alleviate dysfluency in some stutters.

Harrington (1988) proposed that stuttering results from an erroneous internal prediction for the AF of ongoing speech. Harrington started from the assumption that AF is used in the online control of speech articulation to check the timing of the vowel onsets in successive syllables.

Based on a review of previous findings, his theory makes the assumption that the speech system contains a "schedule" of expected time of arrival of the consecutive vowel onsets. If the AF of the onset of a vowel is belated relative to the schedule, the system will interpret it as a temporal error. To correct for this error, the system will delay the end of the consonantal gesture that precedes the vowel, which leads to the repetition or prolongation of the consonant and the entire syllable. This model can naturally explain the detrimental effects of DAF on normal speakers' fluency. Harrington went a step further and hypothesized that PWS stutter because their expectations of the time of arrival of the AF of vowel onsets are *too early*. According to his model, although the cause is rather different, the consequence of this premature expectation of AF is similar to the effect of DAF on normal speakers, that is, the repetition and prolongation of initial consonants. This model of Harringon's has many merits. First, this model offers a straightforward explanation for the phonemenon that dysfluencies in stuttering occur predominantly at the onset of an utterance. Second, it offers a natural and logical explanation for the fluency enhancing effect of DAF on stutterers (see Sect. 1.3.1) because according to this model, DAF offsets the prematurity of the timing expectations and eliminates the erroneous correction attempts that lead to dysfluencies. Thirdly, it can explain the fluency enhancing effect of noise masking based on the assumption that if AF is unavailable, the speech system will stop checking the AF against the timing expectations. In addition, it offers an explanation for the rhythm effect[18]: the regular interval to which syllables are aligned aids the PWS in generating correct expectations of the timing of AF. Finally, unlike other theories of stuttering that focused on AF (e.g., Mysak 1960), Harrington's (1988) theory is capable of explaining the fact that dysfluencies can occur at the beginning (first sounds) of an utterance.

[18] Rhythm effect is a well-known fluency-inducing condition in stuttering. It refers to the following phonemenon: when a PWS voluntarily aligns the speech rhythm (e.g., words or syllables) to a beat or rhythm, there will be a dramatic decrease in the frequency of dysfluency. This phenomenon has been demonstrated for auditorily and visually presented beats, internally generated beats, during oral reading and spontaneous speech (e.g., Hanna and Morris 1977; Stager et al. 1997). A related and similar fluency-inducing condition is choral reading, in which a PWS reads in unison with one or more accompaniers.

Kalveram (1991) proposed a computational model of the control of phonatory and articulatory timing in CV or CVC syllables that focuses on the roles of the so called central pattern generators (CPGs) and of AF. In his model, AF plays the role of checking the timing of the production with that of a planned temporal pattern. This role of AF was active only during stressed syllables (e.g., in German or English), which was a feature of this model that could explain the observations (by the same research group) that the effect of DAF on the timing of the utterance "tatatas" was significant only on the stressed syllable. Kalveram showed that an abnormal activation of the AF-based control during an unstressed syllable could lead to sound repetition-type stuttering at the onset of the ensuing stressed syllable. This design of the stuttering version of their model was also consistent with the empirical findings (again, of the same group) that PWS show DAF effects on not only the stressed syllables, but also the unstressed ones. This model of Kalveram's is similar to the Harrington model reviewed above in the sense that both postulate that 1) AF is an integral part of online timing control in speech production, and 2) abnormalities in such AF-based online timing control lead to stuttering.

Based on more up-to-date knowledge of the neurophysiology of the motor system, Max and colleagues (2004) proposed two possible neural mechanisms of stuttering: 1) unstable and/or under-activated IMs for speech movements and 2) over-reliance on the AF pathway for speech motor control. Their first proposal is rooted in recent motor neuroscientific theories that motor control in general is realized through the usage of inverse and forward IMs (e.g., Wolpert et al. 1995; Kawato 1999; Perkell 1997). In the context of speech motor control, the inverse model is a part of the feedforward pathway that transforms desired acoustic/auditory outputs into vocal-tract configurations and then into articulatory motor commands, whereas the forward model is a part of the feedback pathway that serves to generate predictions of the sensory consequences of speech for comparison with actual sensory feedback. Max and colleagues argued that various empirical findings are consistent with the notion that PWS have defective inverse and/or forward models. As a consequence, they may be unable to consistently generate correct motor commands based on the desired acoustic output, leading to frequent resetting and restarting, which manifest

themselves as moments of dysfluencies. It is also possible, as Max and colleagues suggested, that dysfluencies in stuttering result from erroneous predictions of sensory feedback by defective forward models, leading to unnecessary repair efforts and hence dysfluencies. It should be pointed out that the inverse model-based hypothesis is conceptually similar to the proposal of Neilson and Neilson (1987), and the second, forward model-oriented hypothesis is conceptually reminiscent of Harrington's (1988) theory review above. But Max and colleagues articulated these hypotheses in more general terms and connected more with current neuroscientific knowledge.

The second theory of Max et al. mentioned above is the theory of over-reliance on AF. They hypothesized that the over-reliance may be the consequence of certain defects in the feedforward pathway, potentially consistent with recent findings that white-matter integrity is lower-than-normal in the areas related to oromotor functions (e.g., Sommer et al. 2002; Chang et al. 2008; Watkins et al. 2008; Chang et al. 2011). Due to the intrinsically long latency (~150 ms, see the review in Sect. 1.1.4) in the AF pathway, AF based control can lead to unstable articulatory performance and excessive error, which if too large, can trigger the abortion and restarting of the articulatory unit, leading to moments of dysfluency.

This AF over-reliance theory of Max et al. (2004) was embodied in the detailed DIVA-based computational simulation of Civier et al. (2010). Civier and colleagues modeled the hypothesized over-reliance on AF by using a "stuttering version" of DIVA, which had a feedback control weight (α_{FB}) higher than the "normal" version of DIVA. They showed that under an α_{FB} as high as 0.75 (considerably higher than the normal value 0.15), large deviations from auditory targets (see Sect. 1.2.2 for a review of the DIVA model) can occur. They also modeled a "self-monitoring" unit, not in the original DIVA model, the role of which is to restart an articulatory unit if the AF error exceeds a certain threshold. Unsurprisingly, the incorporation of this self-monitoring leads to behaviors of the model that resemble moments of dysfluencies in PDS. But the dysfluency type in these model simulations is restricted to phoneme or syllable repetition. However, this model does possess some merits in its explanatory power for some fluency enhancing conditions. For example, this model of Civier et al. (2010), rooted in the second

theory of Max et al. (2004), can explain the fluency enhancing effect of masking noise, because noise masking may disengage the self-monitoring system. Also, this model correctly predicts that stutterers will show improved speech fluency under slower speaking rates (see Bloodstein and Ratner 2008, p. 162), because an auditory error accumulate more slowly under slower speaking rates and are therefore less likely to cross the resetting threshold.

Another theory of the neural mechanism of stuttering similar to the internal model-based account of Max and colleagues (2004) has recently been formulated within a new theoretical framework called the State Feedback Control (SFC) model of speech production (Hickok 2011). The SFC model of speech production (Ventura et al. 2009) hypothesizes that a two-way auditory-motor transformation interface, located at the left Sylvian fissure at the junction of the temporal and parietal lobes (Spt), is responsible for 1) generating predictions of the auditory consequences of articulatory commands issued from the primary motor cortex (i.e., forward modeling), and 2) generating corrective motor commands when mismatch between auditory feedback and the predicted auditory feedback arise (a type of inverse modeling). This auditory-motor translation function of the Spt is hypothesized to be "noisy" (p. 417, ibid.) in PWS (see the figure below).

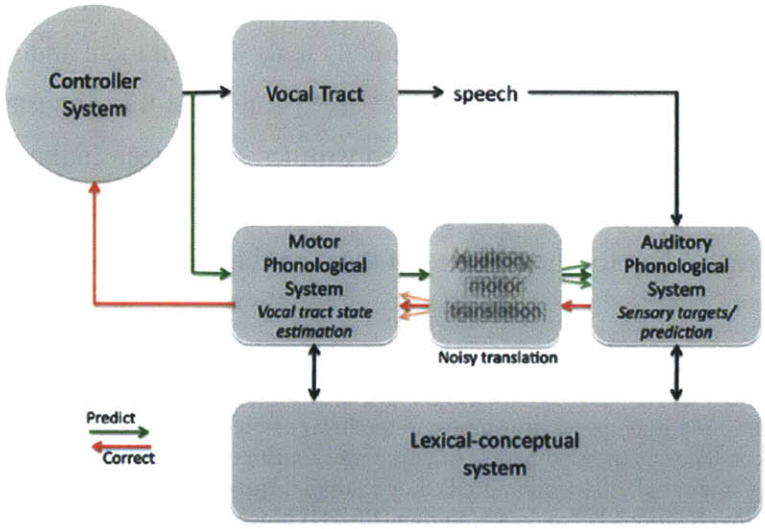


Figure 1.3. A schematic drawing showing the theory of the etiology of stuttering proposed based on the State Feedback Control (SFC) model of speech production (reproduced from Hickok et al. 2011).

There are two consequences of this noisy (or unstable) auditory-motor translation interface. First, erroneous predictions of the auditory consequences of articulatory commands are generated occasionally by the noisy forward model (see the blurred green arrows in Fig. 1.3). These erroneous predictions lead to false alarms of mismatch between predicted auditory consequence and the auditory target of the to-be-produced sound. Second, in an effort to correct for these erroneously predicted auditory mismatches, the noisy inverse model generates invalid corrective motor commands, which may lead to further motor errors[19]. The detrimental effects of these two types processes form a vicious cycle and is manifested as dysfluency events in PWS.

In summary of this section[20], many historical and current theories of the neural mechanisms of stuttering emphasize the sensorimotor interaction (in particular, auditory-motor interaction) in speech production. However, theories differ in their detailed hypotheses regarding which specific components of the speech sensorimotor system are defective and how these defects may lead to moment of stuttering. Whereas some models proposed that impaired internal modeling (Neilson and Neilson 1984; Max et al. 2004, first hypothesis; Hickok et al. 2009) lies at the root of stuttering, others (e.g., Max et al. 2004, second hypothesis; Civier et al. 2010) regard abnormal utilization of AF as the direct cause of stuttering. An important unanswered question is which category of these theories is correct. Many of the premises of these theories are based on somewhat indirect evidence, which is difficult to evaluate and compare (see Chapter 4, Sect. 4.1 for details). The field can benefit considerably from experiments that can directly evaluate the relative merits of these theories, in which the models make unequivocal and mutually exclusive predictions that can be tested directly by the data. In Chapter 4 of this dissertation, we will show that the randomized AF perturbation paradigm constitutes such a test. In that chapter, detailed methods of the experiments and the findings from the PWS and controls will be described, and the implications of these data for the theories of the neural mechanisms of stuttering will be discussed.

---

[19] It should be noted that these hypotheses are similar to the first (defective IM) hypothesis of Max et al. (2004).

[20] There are other modern theories or hypotheses about the mechanism of stuttering, such as the EXPLAN model of Howell (Howell 2004, Civier et al. 2010; submitted). For example, Howell's EXPLAN model posits that stuttering results from abnormal time relations between a planning (PLAN) process and an downstream execution (EX) process, and that auditory feedback plays no direct roles in this EX-PLAN interaction. The EXPLAN model regards the fluency-enhancing effect of auditory masking as a side-effect of the slower speaking rate under auditory masking. In addition, by using the GODIVA model (Bohland et al. 2009), Civier and colleagues (submitted for publication) showed that impaired WM connections that convey efference copies of the motor commands (underlying the ventral primary motor cortex) to the basal-ganglia planning loop can lead to stuttering-like behavior in the simulations of the model, due to insufficient suppression of the currently active syllable representation, which leads to delayed transition into the next syllable. However, because these models do not posit that stuttering results from abnormal interactions between sensory feedback and motor control, they will not be discussed in detail here.

1.4. Summary and aims of the current study

To summarize the preceding review, the interaction between AF and motor control is one of the central topics in speech production research. Previous studies have generated a considerable amount of evidence for an important role of AF in the online control of speech movements. Other empirical observations and theoretical considerations also indicate that this auditory-motor interaction may have important bearings on the etiological mechanism of stuttering. However, due to the focus of previous studies on simple, quasi-static articulation (for the sake of simplicity, and due to technological limitations), we currently still have a very limited understanding of how AF interacts with motor control during the production of speech movement sequences gestures in multisyllabic, connected speech. As a consequence, we lack a carefully verified computational model for sensorimotor interactions during multisyllabic speech. Given that the sequencing of movements and the seamless transitions between them is a defining feature of speech production and that these functions lie at the center of speech disorders such as stuttering, empirical and modeling studies of this sequential speech motor control will be an indispensible step toward a more accurate and thorough understanding of the principles underlying speech motor control in normal and disordered states.

In the current dissertation, we attack this problem from three different angles, which are the themes of following three chapters.

In Chapter 2, we describe a platform that we developed for flexible, subtle and well-controlled manipulation of the AF of F2 during the production of a multisyllabic utterance, and a series of two psychophysical studies performed on normal subjects based on this platform. The results of these experiments generated insights into the spatiotemporal characteristics of auditory-motor integration during multisyllabic articulation.

In Chapter 3, we develop a computational model that theorizes how the normal speech motor system uses AF information to fine-tune the spatiotemporal parameters of articulation in an online, moment-by-moment basis. This model was tuned up by using the experimental data from

Chapter 2 and tested against alternative models. Although this model is not meant to be a comprehensive model of the speech system, it should nonetheless be an important component of future neurocomputational models that encompasses the scope of multisyllabic speech production.

In Chapter 4, a study focusing on the AF-based online speech motor control in stutterers is described. To our knowledge, this is the first systematic investigation of auditory-motor interaction in stuttering. It generated interesting new findings about how the auditory feedback pathway of the speech motor system in a PWS differs from normal and how these differences may related to the neurophysiological bases of this disorder.

Chapter 5 includes a brief recapitulation of the main experimental and modeling results and the new insights generated by this dissertation into the sensorimotor properties of the speech motor system during multisyllabic, connected speech in its normal state and stuttering.

Chapter 2. The role of auditory feedback in the online control of multisyllabic articulation in normal speakers

As reviewed in Sections 1.1 – 1.2, our knowledge of the role of AF in the control of time-varying, multisyllabic articulation remains primitive. Existing empirical evidence based on current methodology suffers from many confounding factors, hence it provides only limited information regarding such a role. Current models of speech motor control do not sufficiently address the interaction between sensory feedback and spatiotemporal aspects of multisyllabic articulatory control. In order to fill this gap, we devised a novel experimental approach, which utilizes real-time digital signal processing to focally and flexibly manipulate the spatial and temporal parameters of auditory feedback while subjects produce a multisyllabic utterance. Two experiments (Experiments 1 and 2) were conducted to separately address the spatial and temporal aspects of articulatory control (described in detail below). Through this unique

approach, we aimed to shed significant new light on the spatiotemporal organization principles of the speech motor system during multisyllabic, connected articulation.

2.1. Experiment 1: The role of auditory feedback in controlling the spatial parameters of multisyllabic articulation.

2.1.1. Methods

2.1.1.1. Subjects

Thirty-six paid subjects (30 males, 6 female; age range: 19.2 – 42.6; median: 24), who were naïve to the purpose and methodology of this study, participated in this experiment. These subjects self-reported to be native speakers of American English and have no history of speech, language or neurological disorders. Nineteen of the subjects were recruited through an emailing list for recruiting research volunteers (*bcs-subjects@mollylab-1.mit.edu*); the remaining 17 were recruited as control subjects for the PWS subjects (See Chapter 4). Since the same experimental protocol was used on the two subsets of subjects, their data were pooled for analysis.

The pure-tone thresholds of each subject were measured with an audiometer (GSI-14, Grason-Stadler, Madison, WI). All these 36 subjects showed thresholds lower than or equal to 20 dB HL at 0.5, 1, 2 and 4 kHz in both ears.

The majority of the subjects were male. This selection bias was based primarily on the consideration that the formant perturbation algorithm used in this experiment (Sect. 2.1.1.3) requires relatively good formant tracking performance, which was typically easier to attain on male voices than female ones (Quatieri 2001). To the best of our knowledge, the only study related to gender difference in the AF-based control of speech production is Chen et al. (2010), which showed slightly greater magnitude (~15%) and longer latency (~13%) of response to perturbation of the auditory feedback of vocal pitch in males than in females. However, to our knowledge, there exists no evidence for qualitative differences in auditory feedback-based articulatory control. Therefore, although caution should be taken when generalizing the

quantitative details in the findings of this study to the general population, the qualitative conclusions of this study should be broadly relevant.

The Institutional Review Board of M.I.T. approved the experimental protocols (Approval 30403000387).

2.1.1.2. Speech task

The subjects read aloud the sentence *I owe you a yo-yo* multiple 160 times. This sentence, which consists of only vowels and semivowels, was chosen to be the stimulus utterance for this study because of two reasons. First, the algorithm used for online perturbation of auditory feedback was designed for perturbing the formant frequency of vowels and vowel-like consonants such as semivowels. Second, the relatively continuous mode of vocal excitation made it possible to track the formant trajectories throughout the entirety of this utterance, which facilitated the extraction of the spatial and temporal measures of the articulation without acquiring simultaneous articulographic data.

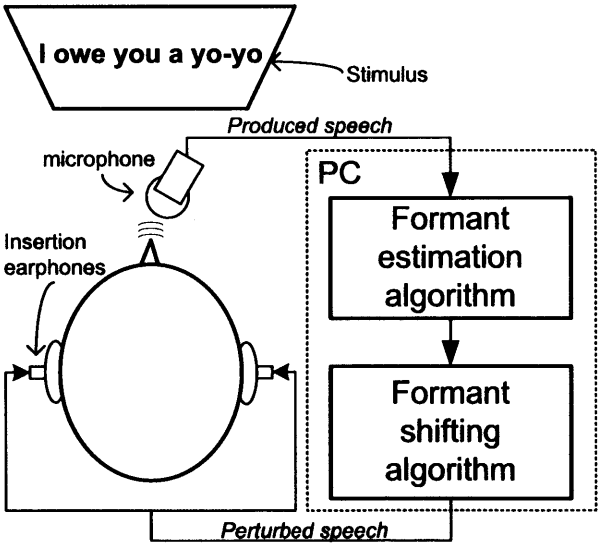


Figure 2.1. A schematic diagram of the experiment setup based on the Audapter platform (Cai et al. 2010).

A schematic diagram of the experiment setup is shown in Figure 2.1. The subject was seated comfortably in front of an LCD monitor, on which the speaking material and instructions were

displayed. A microphone (Audio Acoustica AT-803) was secured at a distance of approximately 10 cm from the mouth of the subject, slightly off the midsagittal plane, with a custom made head-mounted frame. Artificial AF was delivered diotically to the subject through a pair of insertion earphones (E-A-R Tone, Aearo Technologies). The foam ear tips (ER-3A, Etyomtic Research, Elk Grove Village, IL) reduced the level of air-conducted natural AF of speech by 25 – 30 dB (according to the technical specifications). In order to ensure that the artificial AF sufficiently masked the natural AF, the level of the feedback was amplified by 14 dB relative to the sound level measured at the microphone.

To ensure relative consistency of vocal intensity across trials and different subjects, each subject was trained to utter the sentence within a medium range of vocal intensity (74 – 84 dBA SPL) before the data-collection part of the experiment. In addition, the subject were also trained to produce this utterance with a medium speaking rate which corresponded to a sentence duration range of 1 – 1.4 s, so as to ensure a relatively consistent speaking rate between trials and subjects. Following each trial in the data-collection part of the experiment, the subjects were given visual feedback on the screen regarding their success or failure of hitting these target ranges of vocal intensity and speaking rate. On average, the subjects were able to achieve both the intensity and rate target ranges in 91.7% of the trials. However, no trial was discarded from subsequent data analysis on the sole basis of failure to hit one or both of these ranges. In other words, the target ranges of intensity and rate were only used as a means for ensuring the consistency of the manner of speaking, but not used as a criterion for the inclusion of trials for data analysis.

The F2 trajectory of the stimulus utterance contains a series of well-defined local extrema (minima and maxima, Figure 2.2A), which were used as landmarks for locating the phonemes in this utterances. For example, the first local maximum of F2 corresponds to the high-front vowel [i] at the end of the first word "I"; the following local minimum of F2 corresponds to the high-back vowel at the end of the second word "owe"; the next F2 maximum corresponds to the semivowel [j] at the onset of the third word "you", and so on. To facilitate simple and clear

notation, we use a set of short-hands for expressing these minima and maxima, which are summarized in Table 2.1. below.

Table 2.1. Ad hoc phonetic symbols used in the current paper to denote the F2 extrema in the utterance "I owe you a yo-yo".

Symbol	Meaning
[i]	F2 maximum at the end of *I*
$[u]_1$	F2 minimum at the end of *owe*
$[j]_1$	F2 maximum at the onset of *you*
$[u]_2$	F2 minimum at the end of *you*
$[j]_2$	F2 maximum at the onset of the first *yo*
$[u]_3$	F2 minimum at the end of the first *yo*
$[j]_3$	F2 minimum at the onset of the second *yo*

2.1.1.3. Perturbations to the spatial and temporal aspects of formant trajectories

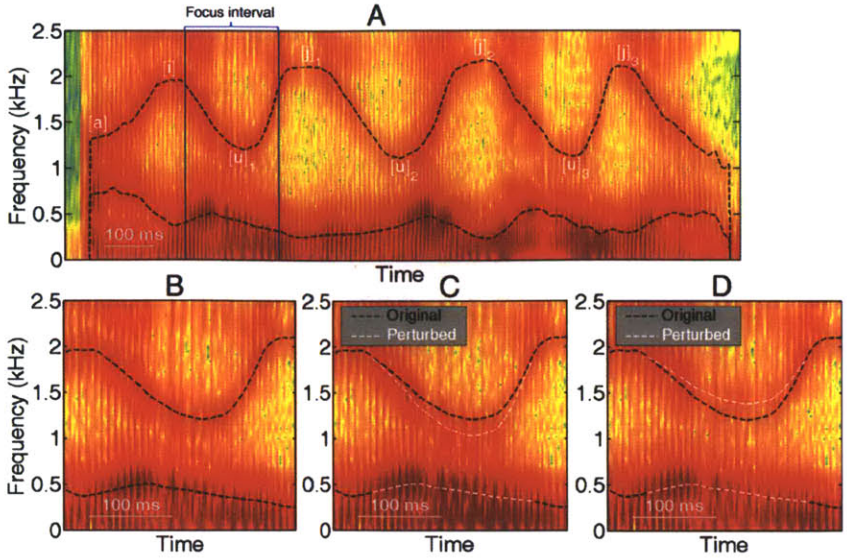


Figure 2.2. Example spectrograms of the stimulus utterance. **A:** an example spectrogram of the stimulus utterance "I owe you a yo-yo" in its entirety. The two dashed black curves show the trajectories of F1 and F2 tracked online with Audapter (see text for details). The time interval bounded by the two vertical lines, referred to as the "focus interval", is the interval in which perturbations to AF occurred. The set of local minima and maxima in the F2 trajectory and their correspondence to the phoneme in this utterance are marked. **B:** a zoomed-in view of the focus interval. **C:** the spectrogram of the Down-perturbed version of this utterance, in the focus interval. The dashed white curve highlights the perturbed F2 trajectory. (The F1 trajectories in the perturbed interval are also shown in white dashed curves, but they are identical to and hence overlap with the unperturbed trajectories, since F1 was not perturbed in this experiment.) The dashed black curves are identical as those in Panel B, to facilitate comparison. **D:** the Up-perturbed version of this utterance. Same format as Panel C.

A MEX[21]-based software, dubbed "Audapter", was used to compute the formant frequencies online when the subjects produced the stimulus utterance. This software was based on previous work at the RLE Speech Communication Group (Villacorta et al. 2007; Boucek 2007; Cai et al. 2010). It uses an autoregression-based linear predictive coding (LPC) algorithm for identifying poles in the sound spectrum, which form a set of candidates for the formants. The sampling frequency of the audio signal was 48 kHz, which was then downsampled to 12 kHz for digital signal processing. For LPC, the orders of 11 and 13 were used for female and male voices,

[21] MEX is an interface between MATLAB and programs written in C++. C++-based programs afford better computational speed than programs written in the native MATLAB environment, which is an interpreted language. This was necessary for the near-real-time constraint of the online formant tracking and perturbation.

respectively. LPC was followed by a dynamic programming procedure that picks real formants (F1, F2, etc.) from these candidates based on a penalty function related to formant values, formant bandwidths and temporal smoothness (Xia and Espy-Wilson 2000).

Unlike traditional methods of manipulating auditory feedback, such as noise masking and DAF, the perturbations to the AF of F1 and F2 used in this study was focal in time. Only the part of the utterance during the second word "owe" and the transition from this word to the beginning of the next word "you" was perturbed. We refer to this time window containing perturbation as the "focus interval" (see Fig. 2.2A). The online detection of this focus interval was based on a set of heuristic rules related to the velocity (1st temporal derivative) of F2. For example, when the F2 velocity turns substantially negative for the first time in the utterance, it is mostly certain that the word "I" has ended and the transition into the high-back vowel [u] in the following word "owe" has begun. Then, when the F2 velocity turns positive, it is mostly certain that the word "owe" has ended and the transition into semivowel [j] at the beginning of the next syllable "you" has started.

Two types of perturbations, which we refer to as the *spatial* and *temporal* perturbations, were used in Experiments 1 and 2, respectively. The spatial perturbation used in Experiment 1 contained two opposite subtypes, namely *Down* and *Up* perturbations, which altered the magnitude of F2 at $[u]_1$, while preserving the timing of the F2 minimum. Examples of the Down and Up perturbations are shown in Figure 2.2.C and D, respectively. The perturbations to the trajectory of F2 were smooth. It gradually ramped up at the beginning of the focus interval and gradually decayed near the end of the focus interval, so that there was no discontinuity in the formant trajectory. This was designed to ensure the subtlety of the AF perturbation and the naturalness of the perturbed F2 trajectory.

These Down and Up perturbations were implemented as a mapping between the original and perturbed F2 values during the focus interval described by the following equation,

$$F_2'(t) = F_2(t) - \Delta F_2(t) = F_2(t) - k \cdot \left(F_2^{\max} - F_2(t)\right), \text{ if } F_2(t) < F_2^{\max}, \quad (2.1)$$

in which F_2 is the original F2, F_2' is the perturbed F2 in the auditory feedback, k is the coefficient of perturbation, set to 0.25 for both Down and Up perturbations for all subjects in Experiment 1, and F_2^{max} is the subject-specific perturbation limit, extracted automatically from the practice trials before the main part of the experiment. According to this Equation, the lower the unperturbed value of F2, the greater the magnitude the perturbation will be, which gives rise to the smooth perturbation pattern as schematically shown in Figure 2.3.

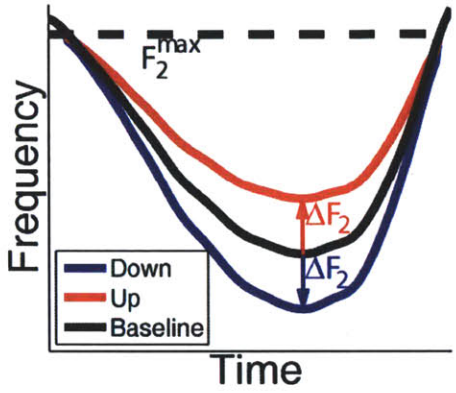


Figure 2.3. Schematic drawings for illustrating the shape of the Spatial (Down and Up) perturbations.

2.1.1.4. Data analysis

The author manually screened all trials from the data-gathering part of each experiment to detect and discarded trials that contain production error including mispronunciations and dysfluencies. In Experiment 1, 0.67% of the trials contained production errors.

The formant tracking algorithm used in the current study is sensitive to quality of the voice. Irregular or amodal glottal cycles, found in breathy speech, as well as excessive nasality, can lead to gross errors in formant estimation, and subsequently, formant perturbation. The author manually identified the trials with gross formant tracking errors and discarded them from further analysis. Trials that were discarded because of formant tracking errors amounted to 4.7% of the total number of trials in Experiments 1.

The F2 trajectories were smoothed with a 17.3-ms Hamming window. An algorithm for identifying local minima and maxima in the F2 tracks was used to extract the timing of the landmarks that correspond to the F2 extrema (Table 2.1) and the magnitude of F2 at those landmarks. As the formant tracks were often noisy and unsmooth, this automatic procedure was not always accurate. The author manually examined all results and manually corrected the results which were deemed problematic. As this manual intervention involved a degree of subjective judgment, it was important to minimize the influence of experimenter bias. To achieve this end, the order of all trials from each experiment was randomized and the experimenter was blinded to the perturbation status (perturbed or unperturbed, and if perturbed, which subtype of perturbation) of the trials.

Statistical analysis involved repeated measures analysis of variance (RM-ANOVA) with subjects treated as a random factor. For a subject, each spatial or temporal measure of the F2 trajectory was averaged across all trials of the same type (e.g., noPert). The within-subject factor perturbation (PERT) took the values of (Baseline, Down, Up). Correction for violation of the sphericity assumption of RM-ANOVA was performed with Huynh-Feldt correction. Post hoc comparisons with control for family-wise errors were conducted by using Tukey's Honestly Significant Differences (HSD).

2.1.2. Results of the spatial perturbation

2.1.2.1. Results from an example subject

Thirty-six subjects were tested under the Down- and Up-type perturbations, which manipulated the magnitude of F2 minimum at $[u]_1$ without changing its timing (see Sect. 2.1.1.3. for details). The F2 trajectories produced by a representative subject (a 19-year old male) under the three perturbation conditions {noPert[22], Down and Up} is shown in Fig. 2.4.

[22] noPert is a shorthand for "no perturbation", which refers to the baseline trials with no perturbation to the formant trajectories.

The F2 trajectories produced by the subject under the baseline (noPert) and Down conditions are shown in Figure 2.4., as the solid back curves and solid blue curves, respectively. The trial-to-trial variation in the shape of the trajectories is obvious from these curves. These two sets of curves are largely overlapping in range. However, careful inspection of the F2 peak at $[j]_1$ reveals that the distribution of F2 at $[j]_1$ in the Down trials is slightly higher than the baseline distribution.

This can be seen more clearly in Figure 2.4.B, which shows the point-by-point average of the F2 trajectories from the noPert and Down trials. In this panel, the effect of the Down perturbation can be clearly seen by comparing the light blue curve with the blue curve. The average F2 trajectories produced under the noPert and Down conditions overlapped almost perfectly from 0 to 200 ms after [i], but started to diverge at the F2 minimum at $[u]_1$. In the transitional period from $[u]_1$ to $[j]_1$, the Down trajectory lay consistently above the noPert trajectory.

The F2 change in the AF caused by the Down perturbation is summarized by the light blue curve in Figure 2.4.C. The darker blue curve in Figure 2.4.C shows the mean difference between the F2 trajectories produced under the Down and noPert conditions. From these two curves in this panel, it can be seen that the compensatory change in the subject's production showed an opposite sign to the AF perturbation. But the compensatory response apparently lagged behind the perturbation by approximately 200 ms.

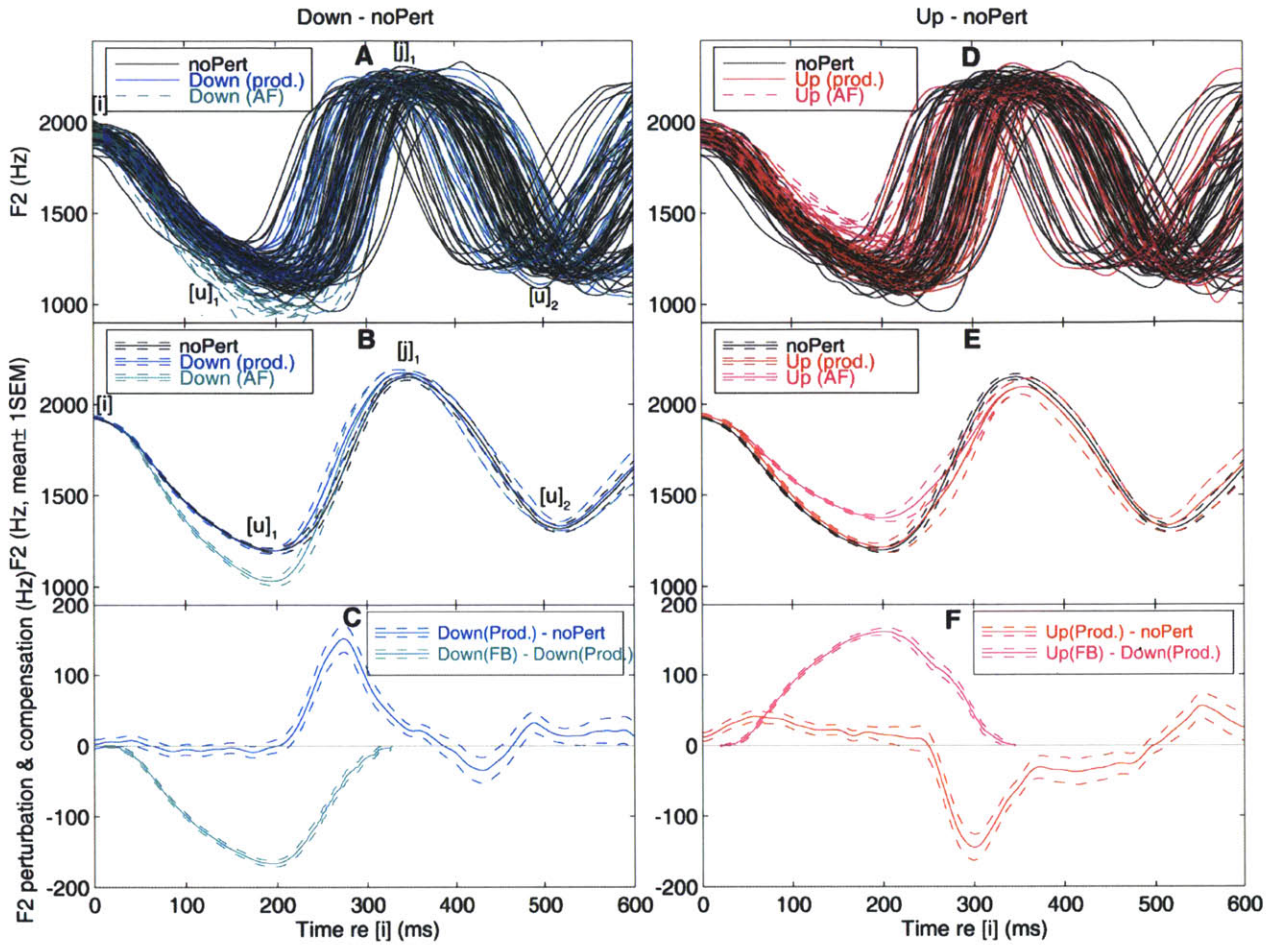


Figure 2.4. Compensatory changes in articulation in response to the Down and Up perturbations in a representative subject. **A.** F2 trajectories from individual trials. The two different conditions: noPert and Down are shown in black and blue, respectively. The lighter blue dashed lines show the perturbed F2 trajectories in the AF in the Down trials. The trajectories are all aligned at the F2 maximum in [i]. **B.** Average F2 trajectories produced by the subject under the noPert (black) and Down (blue) trials. The lighter blue curve show the average F2 trajectory in the perturbed AF in the Down trials. **C**. Blue curve: the mean difference between the F2 trajectories produced under the onPert and Down conditions. Lighter blue curve: the mean time course F2 changes in AF caused by the perturbation. Note that the horizontal axes of Panels A, B and C are aligned so as to facilitate comparison. Panels **D**, **E** and **F** have the same format as Panels A, B and C, respectively, but show the data from the noPert and *Up* trials.

While the average F2 trajectories in Panels B and E of Figure 2.4. provide straightforward and intuitive ways of visualizing the compensatory responses to the perturbations, they suffer from the following potential pitfall: because the local extrema are not perfectly aligned in time due to the natural variation in articulatory timing, the height of the peaks and valleys in the average trajectory may not reflect those in the individual trajectories. For example, the increased

F2 peak at $[j]_1$ in the average trajectory of Down trials (Figure 2.4.B) *could* occur if there was a reduced variation in the timing (i.e., better temporal alignment) of the $[j]_1$ peaks in the individual trials, even if the F2 peak heights didn't change in the individual trials. Similarly, the decreased F2 peak at $[j]_1$ in the average trajectory of the Up trials (Figure 2.4.E) *could* be a result of the increased temporal variation of the $[j]_1$ peak among individual trials, rather than the decreased F2 at the $[j]_1$ peaks in the individual trials. Therefore, to reach solid conclusions about how the spatial measures of the F2 trajectory changed as a function of the perturbation, F2 values need to be extracted from landmarks on *individual* F2 trajectories. For a similar reason, if we need to draw firm conclusions regarding the temporal aspect of the formant trajectory changes, time-interval measures need to be extracted directory from the individual trials as well.

Panels A, C, E, G, I, I and K of Figure 2.5. illustrate schematically the key spatial and temporal dependent variables that were extracted from each single-trial F2 trajectory. As measures of the spatial aspect of articulation, magnitudes of F2 were extracted at four landmark points along the F2 trajectory. These landmarks were based on local F2 extrema during the utterance. These included the F2 minimum at $[u]_1$ (schematically shown in Figure 2.5.A) and the F2 maximum at $[j]_1$ (Figure 2.5.C). In addition, the F2 magnitudes at the temporal *midpoints* between the F2 extrema were also extracted. These included the F2 at the temporal midpoint between $[u]_1$ and $[j]_1$ (dubbed the $[u]_1$-$[j]_1$ midpoint, see the schematic in Figure 2.5.E) and the F2 at the temporal midpoint between $[j]_1$ and $[u]_2$ (called the $[j]_1$-$[u]_2$ midpoint, Figure 2.5.G). These midpoints were examined because with them, we can obtain a more comprehensive characterization of the perturbation-induced change in F2 magnitudes. This was also based on the consideration that the midpoints may afford better statistical sensitivity to F2 magnitude changes because they are less affected by the formant saturation effects (Stevens 1989) at the extrema.

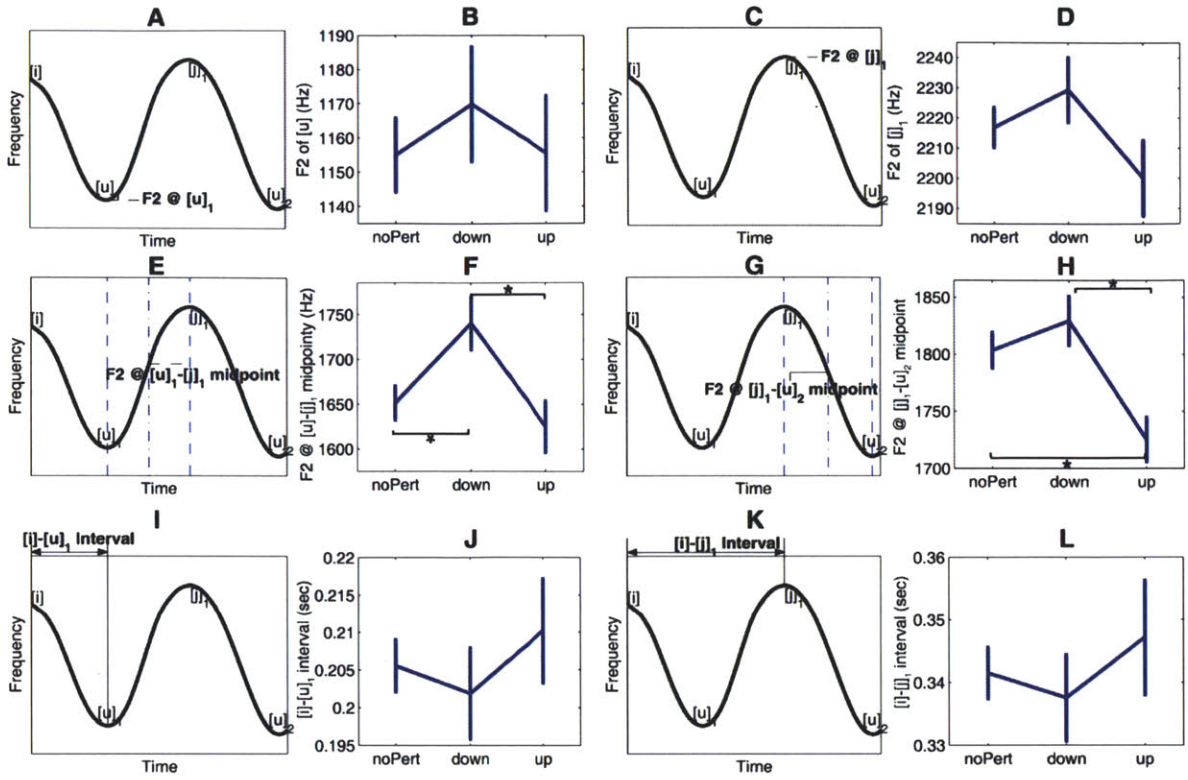


Figure 2.5. Spatial and temporal measures of the F2 trajectory. **A.** A schematic drawing illustrating the extraction of F2 at the trajectory minimum at $[u]_1$. **B.** The F2 at $[u]_1$ produced by the same subject as in Fig. 2.4 under the three perturbation conditions (noPert, Down and Up). The error bars show ±1 SEM around the means. **C.** Schematic illustration of the F2 at the trajectory maximum at $[j]_1$. **D.** F2 at $[j]_1$ produced by the subject under the three conditions. The Down and Up perturbations induced increases and decreases in this F2 maximum, respectively. Due to the large trial-to-trial variability and relatively small number of trials, these within-subject differences were not statistical significant. **E.** Schematic illustration of the F2 value at the temporal midpoint between $[u]_1$ and $[j]_1$, dubbed the $[u]_1$-$[j]_1$ midpoint. **F.** The F2 produced at the $[u]_1$-$[j]_1$ midpoint under the three perturbation conditions. As in Panel D, Increases and decreases in this F2 value can be seen under the Down and Up perturbations, respectively. The difference between the Down and Up conditions reached statistical significance ($p<0.02$, Wilcoxon rank-sum test). So did the difference between the noPert and Down conditions ($p<0.05$). **G**. Schematic illustration of the F2 at the temporal midpoint between $[j]_1$ and $[u]_2$. **H.** The F2 produced at the $[j]_1$-$[u]_2$ midpoint by the subject under the three perturbation conditions. Similar to Panel E and F, the Down and Up perturbations were associated with increased and decreased values of this F2 value, respectively. The difference between the Down and Up conditions was statistically significant ($p<0.005$, Wilcoxon rank-sum test). So was the difference between noPert and Up ($p<0.02$). **I.** A temporal measure of the trajectory: schematic illustration of the definition of the $[i]$-$[u]_1$ interval. **J.** The $[i]$-$[u]_1$ interval produced by the subject under the Down and Up perturbations were on average shorter and longer than the baseline average, respectively. These changes were not significant due to large trial-to-trial variability. **K.** Schematic illustration of the $[i]$-$[j]_1$ interval: another temporal measure. **L.** The $[i]$-$[j]_1$ interval produced by the subject showed shortening and lengthening (relative to the noPert mean) under the Down and Up perturbations, respectively.

As Panels D, F and H of Figure 2.5. show, the mean F2 values at the three landmarks shows a consistent pattern of being higher- and lower-than the baseline (noPert) mean under the Down and Up perturbations, respectively. The mean F2 at the $[u]_1$ minimum only partially showed this trend, which may be attributable to the close proximity between $[u]_1$ and the perturbation onset in time. However, due to the large trial-to-trial variability of F2 (cf. Fig. 2.4A and D), these changes reached statistical significance only at the two midpoint landmarks (Figure 2.5.F and H), but not at the two extrema (Figure 2.5.B and D). However, as will be seen in the next section, since the group-level analysis utilize only the intra-subject mean values of the spatiotemporal measures, the statistical significance of the changes on the intra-subject level is not of the primary concern. Only the group mean values are.

Comparing the patterns of compensatory F2 changes in Figure 2.5. (B, D, F, and H) to the F2 curves shown in Figure 2.4. (Panels B and E), we can see that the directions of change are consistent between each other at the first three landmarks, i.e., $[u]_1$, $[u]_1$-$[j]_1$ midpoint and $[j]_1$, but not at the last landmark, namely the $[j]_1$-$[u]_2$ midpoint. The consistency between these two analysis methods (point-by-point averaging on the time axis and landmark-based data extraction) deteriorates with increasing time from the point of alignment ([i] in Figure 2.4.) because of the accumulating error in temporal alignment with time. This again stresses that the point-by-point trajectory averaging can only be used for first-pass visualization purpose and highlights the necessity of using the landmark-based data extraction for hypothesis testing.

In addition to the spatial measures discussed above, two measures of the articulatory timing were extracted from the F2 trajectory. These included the time spacing between the F2 peak at [i] and the F2 valley at $[u]_1$, called the [i]-$[u]_1$ interval (Figure 2.5.I), and the time spacing between the F2 peaks at [i] and $[j]_1$, referred to as the [i]-$[j]_1$ interval (Figure 2.5.K). Despite being statistically non-significant due to the large trial-to-trial variability, both time intervals exhibited a trend for the mean values to be shortened (i.e., decreased) under the Down perturbation and lengthened (i.e., increased) under the Up perturbation (Figure 2.5.J and L).

The patterns of changes in the F2 magnitudes and the landmark-based time intervals under the Down and Up perturbations seen in this subject are consistent with the hypothesis that the speech motor system is engaged in online monitoring of AF during the production of multisyllabic utterances, and it uses information extracted from AF to adjust both the spatial and temporal parameters of articulation with a short response latency (~150 – 200 ms). However, in order to reach firm conclusion that are generalizable to the general population of normal speakers, we need to examine the data from multiple subjects on the group level.

2.1.2.2. Group Results

After the experiment, the subjects were questioned about whether they were aware of any distortions of the auditory feedback during the experiment. Apart from the higher-than-normal loudness and the differences between hearing one's own voices through natural auditory feedback and through playback or recordings, none of the subjects reported being aware of any deviations of the auditory feedback from the natural pattern.

The inter-subject variability of the patterns of compensatory responses to the Down- and Up-type perturbations can be seen in Figures 2.6. and 2.7., which average F2 perturbation curves and average compensation curves in a subject-by-subject fashion. The results from the Down perturbation are shown in Fig. 2.6. It can be seen that the patterns of the F2 perturbation are relatively consistent across subjects: the mean perturbation curves showed an inverse-bell shape, with the minimum occurring at 150 – 200 ms after the F2 peak during [i]. However, the pattern of compensation varied considerably across subjects. While some subjects exhibited a average pattern of compensation consisting prominent peak of F2 increase (re noPert) at 100 – 200 ms following the peak perturbation (e.g., subjects PFS_S01, 04, 06, 08, 10-13, 18, 20, 22, 25, 29, 31, 33, 36), other subjects produced compensatory response that were more complicated in shape. For example, in subjects 02 and 09, there are two peaks of F2 increase following the peak

perturbation; in subject 05, 16 and 34, the peak of F2 increase were preceded by almost equally prominent F2 decreases.

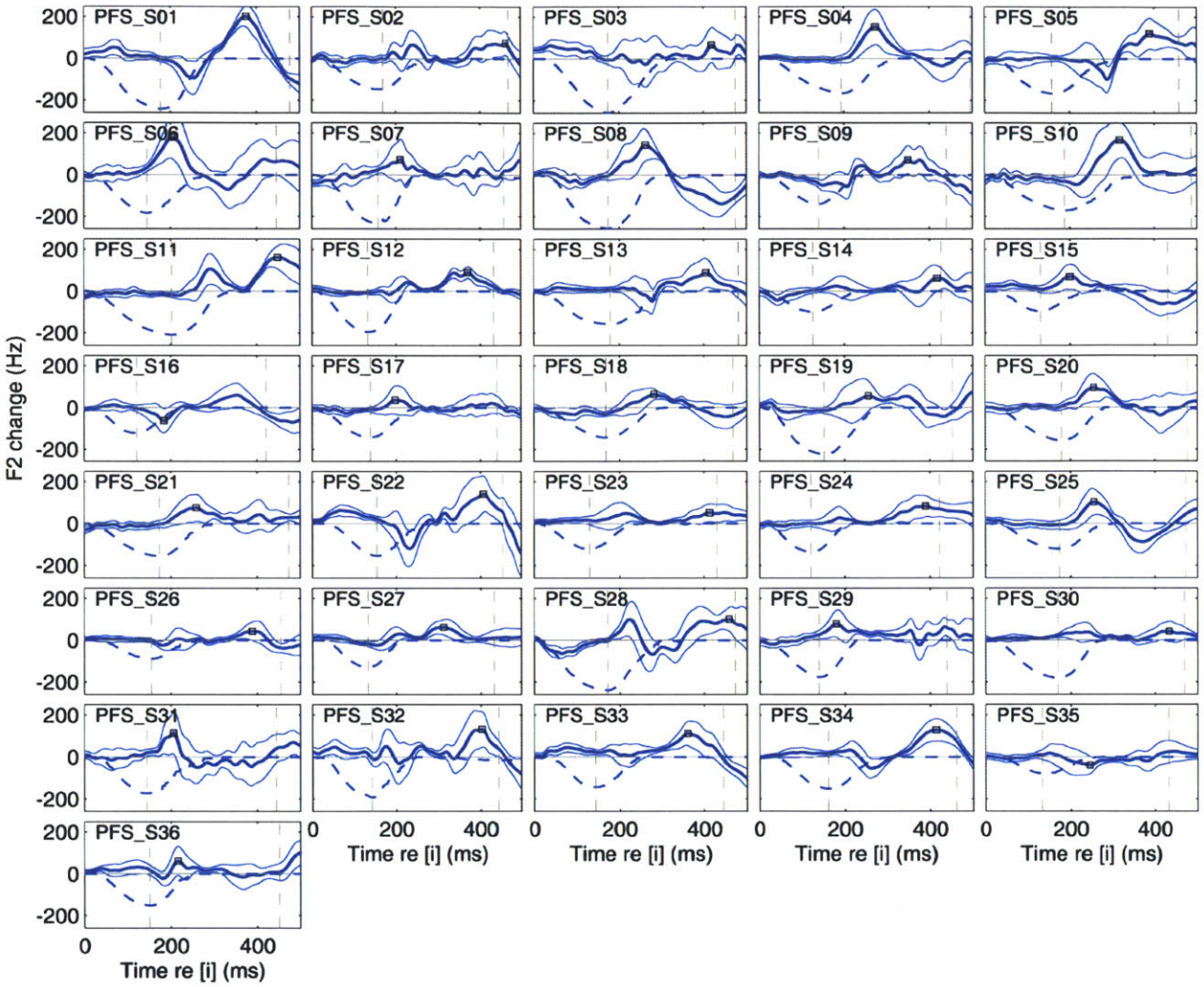


Figure 2.6. The mean F2 perturbation profile of the Down perturbation and the subject's compensatory changes in F2 in their productions. The data from all 36 subjects in Experiment 1 are shown here. Each panel corresponds to an individual subject. In each panel, the dashed line shows the average change in the AF of F2 due to the Down perturbation in the Down-perturbed trials. The thick solid curve shows the mean difference between the mean F2 trajectories produced under the Down and noPert conditions. The thinner solid curves show ±1 SEM around the mean. Before averaging, all F2 trajectories were aligned at the F2 maximum at [i]. The two dashed vertical lines indicate the time interval between 0 and 300 ms after the maximum perturbation point. The black square indicates the local extrema (i.e., minima and maxima) that has the greatest absolute value in this time interval. This is the point at which the peak compensation under Down perturbation for the corresponding subject will be computed (see Fig. 2.8.).

Similar observations can be made from the mean perturbation and compensation patterns under the Up perturbation, which are shown in Fig. 2.7. The F2 changes in AF due to the

perturbation, as indicated by the dashed curves, typically peak at about 150 – 200 ms following the F2 peak of [i]. In response to this perturbation, the majority of the subjects (PFS_S01, 03, 04, 07, 10-13, 15, 17, 18, 21, 23, 24, 25, 31, 33) showed a prominent downward F2 within 300 ms following the peak of perturbation. However, this pattern could not be observed in all 36 subjects.

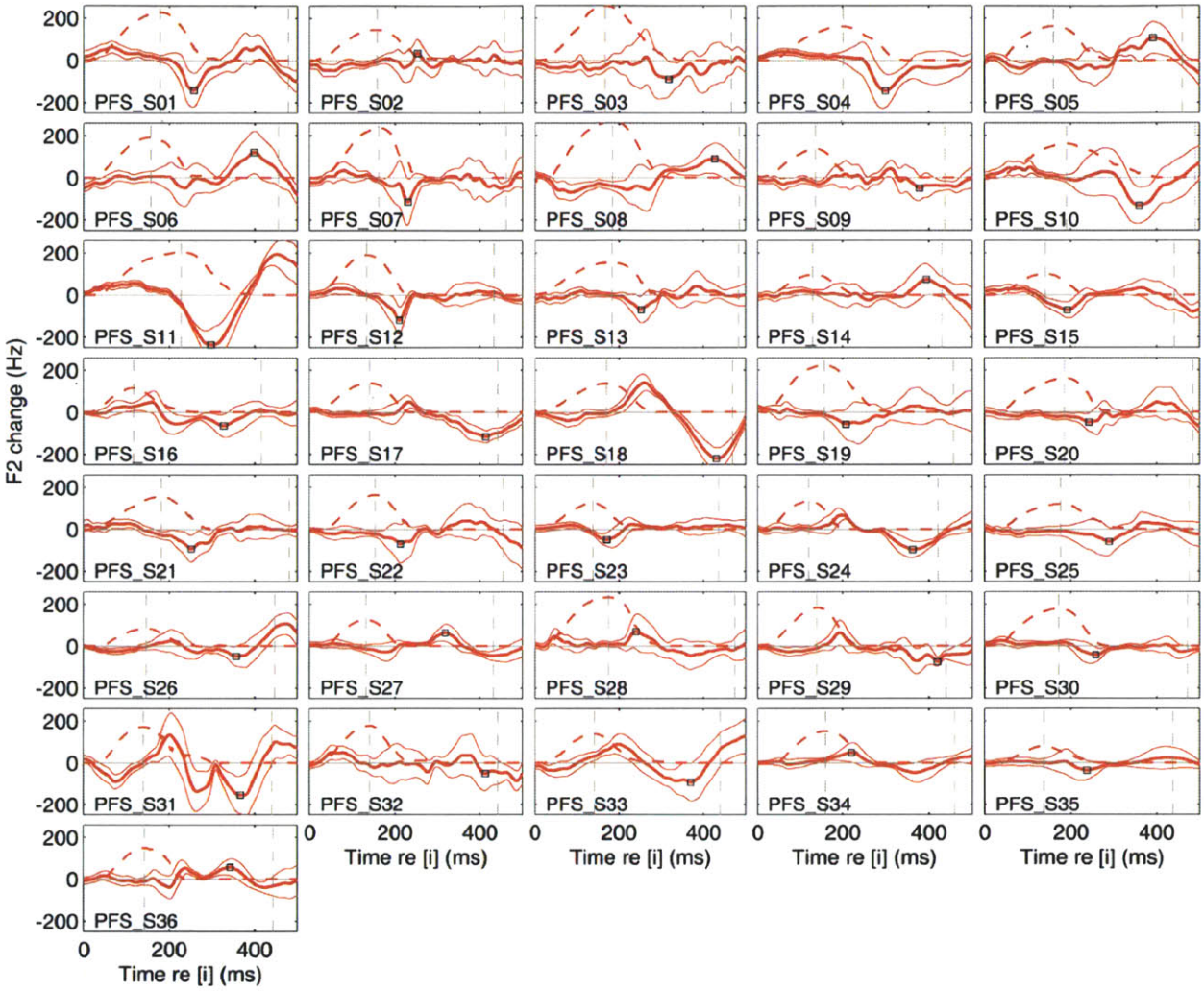


Figure 2.7. The mean F2 perturbation profile of the Up perturbation and the subject's compensatory changes in F2 in their productions. The format of this figure is the same as that of Fig. 2.6.

In order to quantify the direction and amount of compensation in an objective way, we extracted *a peak of compensation* from each subject's compensation curve according in the following way. Within the time window from 0 to 300 ms after the extremum in the mean perturbation curve (i.e., the dashed curves in Figs. 2.6. and 2.7.), the local extremum in the compensation curve (i.e., the solid curves in Figs. 2.6. and 2.7.) with the greatest absolute value

was defined as the peak of compensation. It should be noted that since this simple definition doesn't incorporate a priori assumption about the direction of compensation, it is not biased to favor the direction of compensation or the direction of following responses.

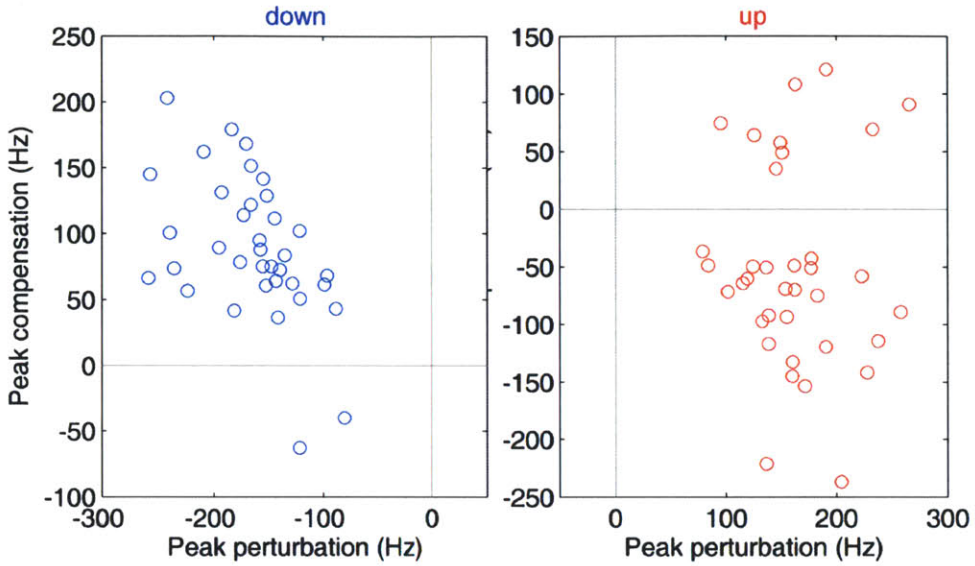


Figure 2.8. The relationships between peak perturbation and peak compensation under the Down (Left) and Up (Right) perturbations. In the left panel, positive values on the vertical axis correspond to compensatory responses (i.e., production changes in the direction opposite to the perturbation); in the right panel, negative values on the vertical axis correspond to compensatory responses. In each panel, every data point corresponds to one individual subject.

The small black squares in Figs. 2.6. and 2.7. show the peaks of compensation extracted using this definition. The magnitudes and signs of these peaks, as well as their relations with the extreme amount of perturbation, are summarized in Fig. 2.8. In the left panel of Fig. 2.8., it can be seen that all but two subjects showed positive peaks of compensation in response to the Down perturbation. From the right panel the same figure, we can see in that 27 of the 36 subjects, the peaks of compensation were associated with negative F2 changes under the Up perturbation, which were also in the direction opposite to the perturbation-induced F2 changes under the Up perturbation.

For each subject, we defined a measure called *ratio of compensation* as the ratio between the magnitude of the afore-defined peak of compensation and the magnitude of maximum perturbation, corrected for the signs in a way such that positive values of this ratio corresponded

to compensatory articulatory changes and negative ones corresponded to articulatory changes that followed the direction of perturbation. Figure 2.9. shows the Tukey's box-plots of the ratios of compensation under the Down and Up compensations. The ratios of compensation were significantly greater than zero under the Down (Two-tailed one-sample t-test: $t_{35}=9.57$, $p<0.0001$) and Up ($t_{35}=3.71$, $p<0.001$) perturbations. On average, the ratios of compensation were 52.7%±5.5% and 33.6%±9.1% (arithmetic mean±1 SEM) under the Down and Up perturbation, respectively. In other words, the subjects' compensatory responses amounted to approximately 53% and 34% of the perturbations under the Down and Up perturbation, respectively.

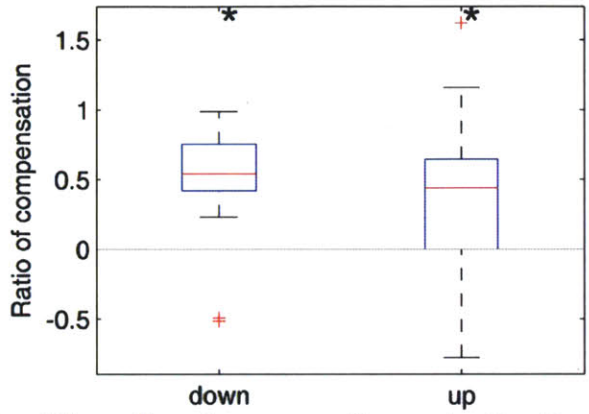


Figure 2.9. Box-plots of the ratios of compensation under the Down and Up perturbations. The ratio of compensation is defined as the ratio between peak compensation and peak perturbation, corrected for the sign of F2 change. Positive values correspond to compensatory response. A value of 1 corresponds to full compensation. The asterisks indicate significant difference from zero ($p<0.0001$ for Down and $p<0.001$ for Up, two-tailed one-sample t-test; see text for details).

To obtain the group-average pattern of compensation, the perturbation and compensation curves from Figs. 2.6. and 2.7. are pooled across subjects, aligned (see Fig. 2.10.A and B) and averaged point-by-point along the un-normalized time axis. The resultant group-average perturbation and compensatory curves are shown by the dashed and solid curves in Fig. 2.10.C. The peaks of compensatory responses are misaligned considerably between subjects under both the Down and Up perturbations. As a result, the shapes and peak magnitudes of the average compensatory curves (solid curves in Fig. 2.10.C) are vastly different in shape and magnitude from the compensation curves from the individual subjects.

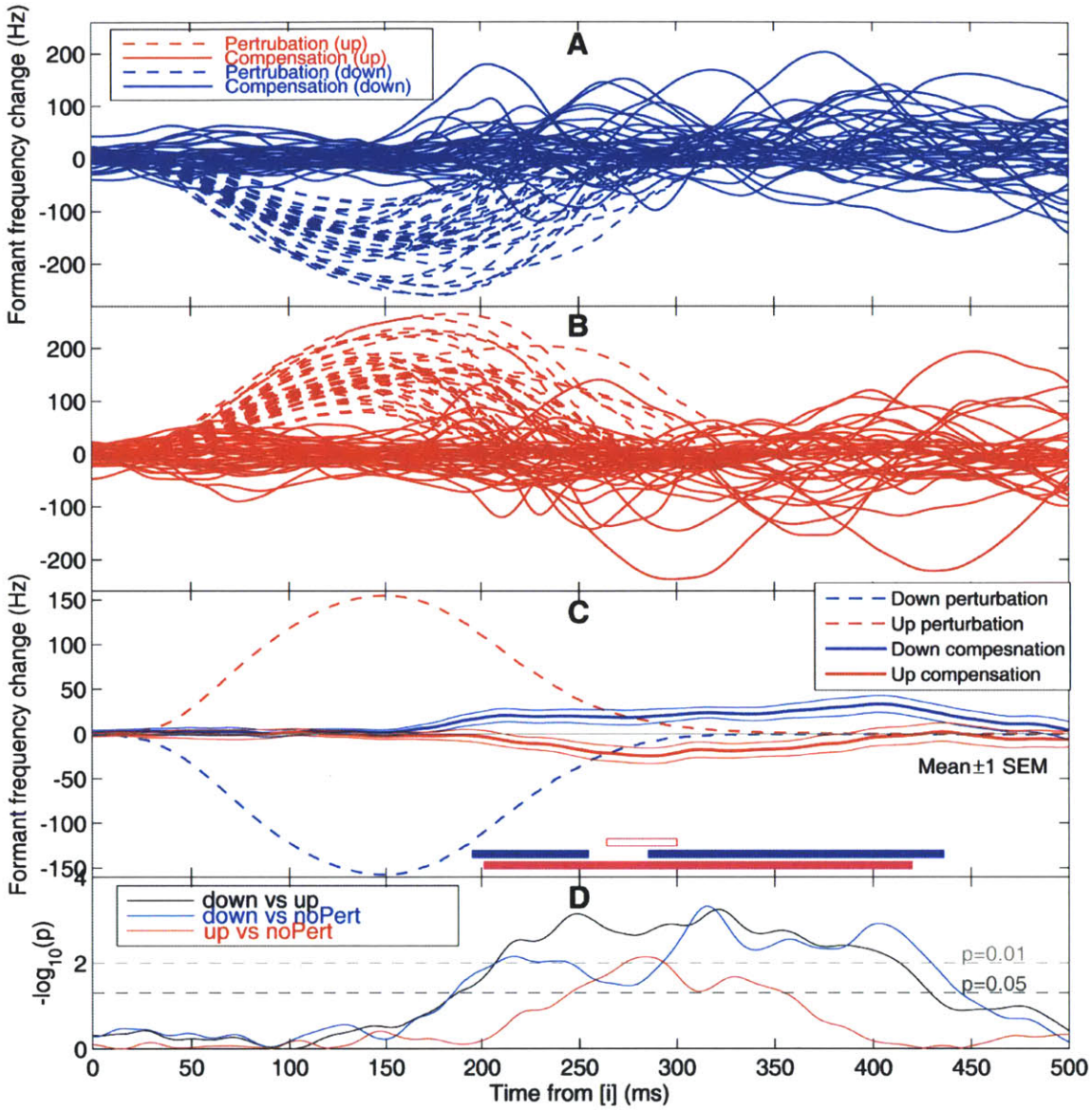


Figure 2.10. Compensatory articulatory adjustments at the group level. **A.** Average temporal profiles of the changes in the produced F2 trajectory (from the average noPert baseline) under the Down perturbations in the 36 individual subjects. Each solid curve corresponds to the articulatory compensation of one subject. The dashed curves show the average temporal profiles of the perturbations to F2. B. The same format as A, but for Up perturbation. **C.** Average perturbation (dashed) and compensation (solid) profiles across the 36 subjects. The thin curves around the compensation curves show ±1 SEM across the subjects. The horizontal magenta bar at the bottom of the panel indicates the time interval in which the difference between the Down and Up compensation curves reached significance at FDR=0.05 (two-tailed t-test). The solid blue bars indicate the time intervals containing significant difference between the Down and noPert conditions (FDR=0.05). The unfilled red bar shows the time interval containing significant difference between the Up and noPert condition (uncorrected p=0.02). **D.** Negative 10-based logarithm of the p-values of three statistical comparisons: Down vs. Up (black, matched two-sample t-test), Down vs. noPert (blue, one-sample t-test), and Up vs. noPert (red, one-sample t-test). Note that higher values correspond to greater statistical significance. The two uncorrected thresholds p=0.01 and p=0.05 are indicated by the two horizontal dashed lines.

The thin solid lines surrounding the group-average compensation curves in Fig. 2.10.C show ±1 SEM across the 36 subjects. The two compensation curves are both close to the zero line and not separated from each other substantially in early parts of the perturbations. However, starting at approximately 150 ms after the onset of the perturbation, the two compensation curves begin to diverge from zero and from each other. The black curve in Fig. 2.10.D shows the p-values from the paired t-tests (df = 35) between the Down and Up compensation curves. The p-value was non-significant (close to unity) until around 200 ms after [i], when it quickly broke through the 0.05 (uncorrected) threshold and then the 0.01 threshold at approximately 210 ms after [i]. With an $\alpha = 0.05$, the significant difference between the two sets of compensation curves at the 0.05 level lasted until approximately 420 ms after [i].

While the black curve shows the significance of the separation between the compensation curves from the Down and Up conditions, it doesn't contain direct information about the significance of the F2 production changes under the individual perturbation types. This information is shown by the blue and red curves in Fig. 2.10.D for the Down and Up perturbations, respectively. Unsurprisingly, the levels of significance for the individual perturbation types were lower than that of the Down-Up comparison. In fact, each curve broke the $p = 0.01$ threshold for only brief moments of period.

Whereas Panel D of Fig. 2.10. shows results from uncorrected statistical comparisons, the results of comparisons under False Discovery Rate (FDR, Benjamini and Hochberg 1995) are shown by the solid horizontal bars in the bottom part of Panel C. The Down-Up and Down-noPert comparisons both showed time intervals with significant difference under the FDR correction. However, no time window with significant corrected difference was found under the Up-noPert comparison.

As discussed in the previous section, the un-normalized time axis used in Fig. 2.10. is suitable for a first-pass examination of the data and for estimating the latency of compensation on the group level, but it suffers from two shortcomings: 1) it doesn't correct for the misalignment in time of the F2 extrema across trials and subjects, which may lead to unwanted smoothing of the pattern of compensation; and 2) it intermingles the F2 changes due to timing and magnitude (spatial) adjustments. In order to isolate the spatial adjustments from the timing ones, the time axis was normalized in a piecewise linear fashion. The F2 trajectories from individual trials were anchored at the set of F2-extremum landmarks in Table 1 ([i], $[u]_1$, $[j]_1$, $[u]_2$, $[j]_2$, $[u]_3$ and $[j]_3$); the F2 trajectories between adjacent landmarks were computed through linear interpolation in time. This piecewise normalization isolates compensatory corrections in the magnitude of F2 from the adjustment of the timing of the F2-extremum landmarks.

Figure 2.11.A shows the average F2 magnitude changes under the Down and Up perturbations along the piecewise-normalized time axis. The difference between the Down and Up conditions was statistically significant within a time interval between $[u]_1$ and $[u]_2$ (FDR = 0.05, see the magenta bar in Fig. 2.11.A). If the gradual buildup to the significant differences and the subsequent decay are included, the magnitude compensation spanned a longer time interval, from $[u]_1$ to $[j]_2$. The largest F2 magnitude adjustments are seen near the temporal midpoints between $[u]_1$ and $[j]_1$ and between $[j]_1$ and $[u]_2$. Interestingly, the compensation magnitude shows a "dip" near the $[j]_1$, an F2 maximum.The reason for this decreased F2 compensation magnitude around the semivowel is unclear, but may be related to a nonlinear saturation relation between articulatory position and formant frequency for this phoneme (Stevens 1998). When the F2 changes were analyzed at individual landmark points, significant compensatory changes were also observed. These landmarks included the F2 minimum at $[u]_1$, the temporal mid-point between $[u]_1$ and $[j]_1$, the F2 maximum at $[j]_1$, and the temporal mid-point $[j]_1$ and $[u]_2$ (Fig. 2.11.B-E). At each of these landmarks, RM-ANOVA indicated a significant main effect by perturbation condition (noPert, Down and Up, $F_{2,70}$=5.49, 11.12, 14.88, and 12.77, with $p<0.01$, 0.0001, 0.000001, 0.0001 for the four above mentioned landmarks, respectively). Pair-wise

Tukey's HSD comparisons between the Down and Up conditions reached significance for all three landmarks as well ($p<0.05$ corrected for all landmarks).

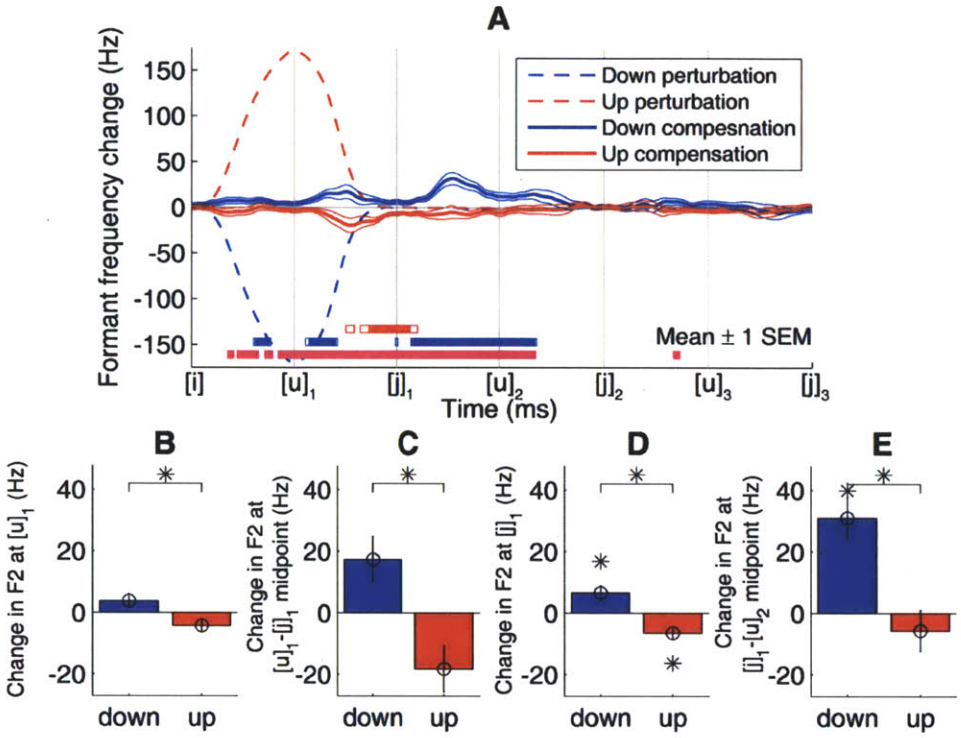
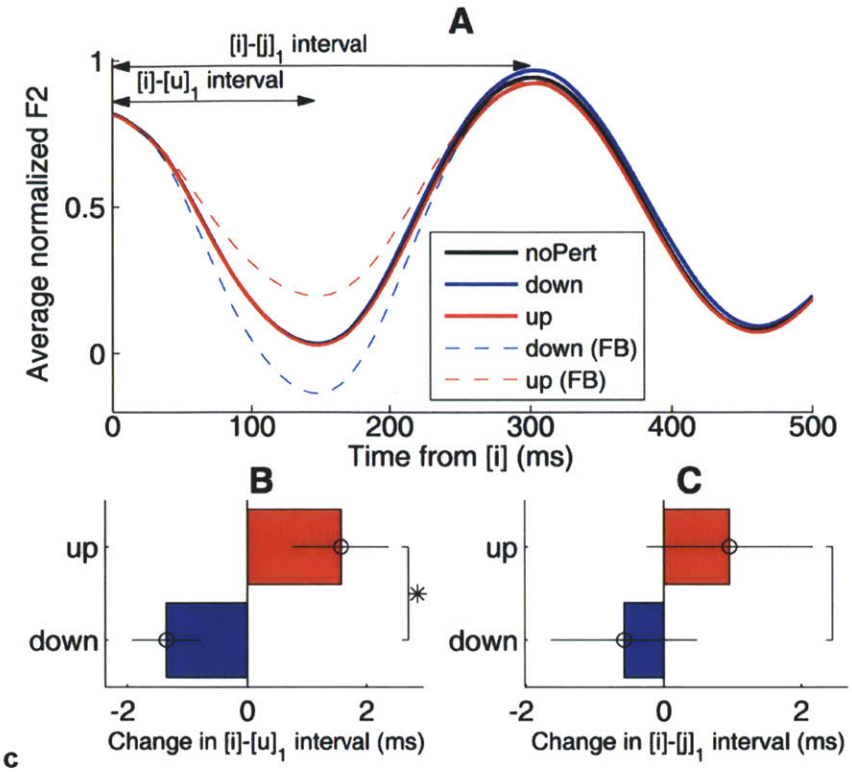


Figure 2.11. Compensations in the spatial parameters of articulation in response to the Down and Up perturbations. **A.** Group-mean perturbation and compensation profiles plotted on the piecewise normalized time axis. The megenta bar near the time axis indicates the intervals of significant difference between the responses to Down and Up perturbation under paired t-tests. Unfilled portions: $p<0.05$ uncorrected; filled portions: corrected at FDR=0.05. **B**: changes in the value of F2 at the minimum in $[u]_1$. **C**: change in value of F2 at the temporal midpoint between the F2 minimum in $[u]_1$ and the F2 maximum in $[j]_1$. **D**: change at the F2 maximum in $[j]_1$. **E**: change at the midpoint between the F2 maximum in $[j]_1$ and the F2 minimum in $[u]_2$ in "you". Error bars: ±1 SEM. Asterisks show significant difference at $p<0.05$ (post hoc Tukey's HSD following RM-ANOVA). Note that the y-scales are identical in Panels B, C, D and E.

In addition to these changes in the magnitude of F2, which reflected feedback-based control of the spatial parameters of articulation, we also observed significant changes in the timing parameters of the F2 trajectory under the auditory perturbations. The perturbations conditions significantly affected the subjects' produced [i]-$[u]_1$ interval, i.e., the interval between the F2 maximum at [i] and the F2 minimum at $[u]_1$ ($F_{2,70}=5.92$, $p<0.005$) and this interval was significantly different between the Down and Up conditions ($p<0.05$ corrected, post hoc Tukey's HSD). On average, this interval shortened and lengthened under the Down and Up perturbations,

respectively (Fig. 2.12.B). If the F2 minimum at $[u]_I$ is defined as the end time of the syllable "owe", this observation indicates that the Down and Up perturbations led to an earlier- and later-than-baseline termination of this syllable, respectively. In other words, these perturbations altered the articulatory timing *within* this syllable. In comparison, the [i]-$[j]_I$ interval, namely the interval between [i] and $[j]_I$, exhibited a similar, but non-significant trend of change ($F_{2,70}$=0.73, p>0.45, Fig. 2.12.C). Therefore, if the F2 maximum at $[j]_I$ is regarded as the onset of the syllable [ju] ("you"), it can be seen that the Down and Up perturbations didn't significantly alter the onset timing of this following syllable (i.e., *between-syllable* timing).

To summarize the findings of Experiment 1, considerable between-subject variability exists in the compensatory response to the Down and Up perturbation. However, a consistent and statistically significant pattern of spatiotemporal compensation did emerge at the group level. On average, subjects responded to the Down and Up perturbation by altering the values of F2 in their productions starting approximately 180 ms following the onset of the perturbation. The group-average F2 magnitude adjustments were in the directions opposite to the perturbation. These compensatory F2 adjustments lasted for a time window longer than the perturbation itself. These findings are evidence for the involvement of auditory feedback in the online feedback-based guidance of the spatial aspect of multisyllabic articulation. As for the role of auditory feedback in controlling articulatory timing, such a role was observed only in the control of within-syllable timing (Fig. 2.12.B), but not in the control of between-onset timing (Fig. 2.12.C). These differences were very small. There are two possible explanations for this pattern: 1) the syllable onset times may be highly pre-programmed (e.g., Fowler 1980), to the extent that changes in auditory or other sensory feedback states are not capable of affecting the syllable-onset times; and 2) auditory feedback is utilized by the speech motor system in the online control of syllable timing, but the Down and Up perturbations used in Experiment 1 might not be the most appropriate types of perturbation to demonstrate such a role of auditory feedback.



c

Figure 2.12. Timing adjustments under the Down and Up perturbations. **A**: Grand average F2 trajectories aligned at the F2 maximum in [i] of "I". The time axis shows un-normalized (real) time and includes only an early part of the utterance, from [i] to $[u]_2$. Error bands are omitted for clarity of visualization. F2 was amplitude normalized before averaging across subjects. The doubled-sided arrows schematically indicate the time-intervals of which the compensatory changes are summarized in Panels B and C. **B**: change in the $[i]\text{-}[u]_1$ time interval. **C**: change in the $[i]\text{-}[j]_1$ interval. Asterisks: significant difference at $p<0.05$ (post hoc Tukey's HSD following RM-ANOVA). Note that the y-scales are different between the plots.

2.2. Experiment 2: The role of auditory feedback in controlling the temporal parameters of multisyllabic articulation.

In order to address the unanswered question about the role of AF in the online control of intersyllabic timing during multisyllabic articulation, we devised two new types of perturbations of F2 trajectories, namely temporal perturbations. Unlike the spatial perturbations used in Experiment 1, these temporal perturbations alter the timing of the F2 minimum associated with $[u]_1$ in the subjects' AF. We hypothesized that with these new perturbations, significant changes

in the subjects' articulatory timing would be observed, which would support a role of auditory feedback in the online control of both within-syllable and between-syllable timing.

2.2.1. Methods

Twenty-eight subjects (24 male, 4 female; age range: 19.2 – 47.1, median: 24.7) participated in this experiment. Seventeen of the 24 subjects also participated in Experiment 1 (Sect. 2.1.). Eight of the 24 subjects also participated in the spatial perturbation experiment (i.e., Experiment 1); these subjects were recruited as controls for the PWS in the project on stuttering (See Chapter 4). These eight subjects were tested under the spatial and temporal perturbation in a randomized and counterbalanced order.

The design of Experiment 2 was similar to that of Experiment 1. Each experiment consisted of 20 blocks of eight trials which each contain two perturbed (one Accel and one Decel) trials and six noPert (baseline) trials. The order of the trials within each block was randomized with the constraint that no two consecutive trials both contain perturbation. Trials with speech errors and/or dysfluencies, which amounted to 0.80% of all trials in Experiment 2, were excluded from further analysis. The experimenter, blinded from the perturbation status of all trials, manually examined the quality of formant tracking and perturbation and discarded those trials which contain gross formant tracking or perturbation failure. Such trials amounted to 3.6% of all trials.

In the statistical analysis of the data, the within-subject factor perturbation type took the values of (Baseline, Accel, Decel). The rest of the statistical procedures were the same as in Experiment 1.

Unlike the spatial perturbations used in Experiment 1, the temporal perturbations used in Experiment 2 manipulated the timing, rather than the magnitude of the F2 minimum $[u]_I$. Two opposing subtypes of temporal perturbation, namely Accelerating (Accel) and Decelerating (Decel) perturbations, led to earlier- and later-than-baseline occurrence of the F2 minimum $[u]_I$

in the AF, respectively. As the examples in Fig. 2.13.C show, the Accel perturbation altered the F2 trajectory in such a way that the $[u]_1$ occurs earlier than the actual timing of this F2 minimum. This effectively reduced the duration of the syllable [ou] and elongated the transition from the end of [ou] to the beginning of the following syllable, [ju]. The Decel perturbation had the opposite effect: it delayed the timing of the $[u]_1$ (see example in Fig. 2.13.D).

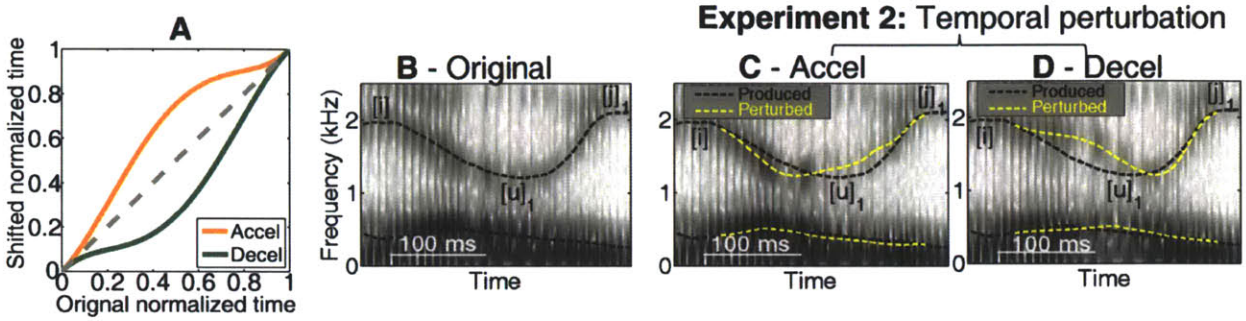


Figure 2.13. The temporal (Accel and Decel) perturbation used in Experiment 2. **A.** Schematic drawings illustrating the mathematical details of temporal (Accel and Decel) perturbations. The time-warping functions used in the temporal (Accel and Decel) perturbations. **B**: the spectrogram of an original (unperturbed) recording during the focus interval, for comparison with the perturbed spectrograms in Panels C and D. **C** and **D:** the Accel- and Decel-perturbed versions of the same utterance. The yellow dashed curves indicate the perturbed formant trajectories. Notice that unlike the Down and Up perturbations, the Accel and Decel perturbations manipulated both F1 and F2.

The Accel and Decel perturbations were achieved through time-warping in the focus interval. The time-warping was governed by the following equation,

$$F_2'(\frac{t-t_0}{T_{est}}) = F_2(W(\frac{t-t_0}{T_{est}})), \text{ when } t < t_0 + \overline{D}, \qquad (2.2)$$

wherein t_0 is earliest time at which $F_2(t) < F_2^{\max}$ is satisfied (i.e., onset of the focus interval). T_{est} is the estimated duration of the focus interval, updated online based on the preceding trials, $W(\bullet)$ is a 4th-order polynomial time-warping function shown in Fig. 2.13.A, and $\overline{D}$ is the subject-specific average duration of the focus interval computed from previous trials, which was

updated adaptively during the course of the experiment. Perturbations to the trajectory of the first formant (F1) were done in a similar manner.

The time-warping function $W(\bullet)$ took different forms for the Decel and Accel perturbations. In Fig. 2.13.A, the green curve in Fig. 2.13.A shows the time-delaying warping used in the Decel perturbation; the orange curve in the same panel shows the time-advancing function used for the Accel perturbation. The time-warping in the Accel perturbation was non-causal and hence required predictions of future F1 and F2 values. This prediction was achieved by using average F1 and F2 trajectories during the focus intervals of previous trials. Due to the naturally occurring trial-to-trial variation in the magnitude of the F2 minimum, a certain amount of mismatch in the value of the F2 minimum at $[u]_1$ between the perturbed auditory feedback and the production were inevitable in the Accel perturbation. Figure 2.14.A summarizes the mismatch in individual subjects and on the group level. It can be seen that although for the individual subjects, some trials contained relatively large (~100 Hz) error in the prediction of the F2 minimum at $[u]_1$, on the group level, the average prediction error was relatively close to zero. The matching error for the F2 minimum was -3.43±2.88 Hz for the 28 subjects, which was not statistically significantly different from zero (t_{27}=-1.19, $p>0.24$).

It should be noted that Equation (2.2), which governs the temporal perturbations, did not specify explicitly the amount of change in the timing of $[u]_1$. The amount of $[u]_1$ time shift depends on the shape of the F2 trajectory in the focus interval, which varied from trial to trial and from subject to subject. Figure 2.14B summarize of the $[u]_1$ timing shifts across under the Accel and Decel conditions for the 28 subjects. It can be seen that on average, the Accel perturbation led to a 43.89-ms advancing of $[u]_1$ in time, while the Decel perturbation led to a 24.21-ms delay of $[u]_1$.

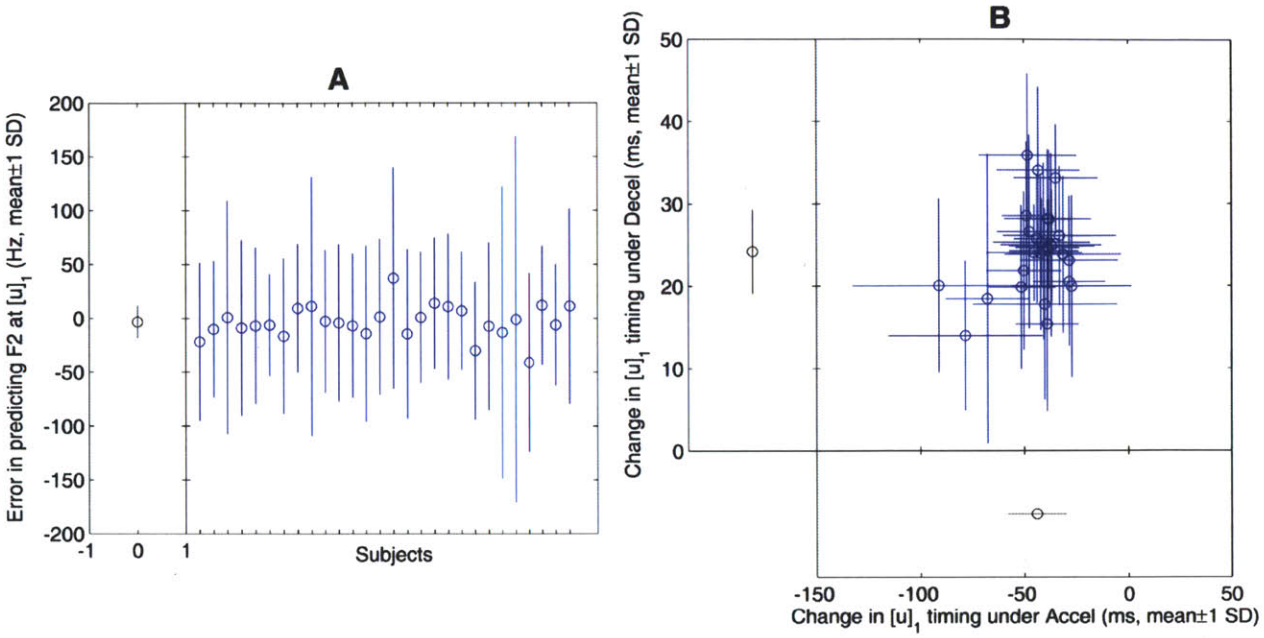


Figure 2.14. Summary statistics of the spatiotemporal changes in the AF due to the temporal perturbations. **A.** Errors in the prediction of the magnitude of the F2 minimum at $[u]_1$ under the non-causal Accel perturbation. Right: mean±1 SD of the mismatch of the F2 at $[u]_1$ between the perturbed and actual F2 trajectory across all Accel trials in each subject. Left: mean±1 SD of the $[u]_1$ F2 mismatch across the 11 subjects. **B.** Change in the timing of the F2 minimum $[u]_1$ in the Accel and Decel perturbations. The $[u]_1$ timing change under the Accel and Decel perturbations are shown on the horizontal and vertical axes, respectively. Negative values indicate advances in time and positive values indicate delays. In the top-right panel, the 28 circles correspond to the 28 subjects. Associated with each data point are the horizontal and vertical bars show ±1 SD across all Accel and Decel trials in the subject. The bottom panel shows mean±1 SD of the perturbation-induced change in the timing of $[u]_1$ across subjects under Accel perturbation; the left panel under Decel perturbation.

2.2.2. Results of the temporal perturbation

After the completion of the experiments, subjects were asked whether they were aware of any distortion of the auditory feedback. Six of the 28 subjects (21.4%, higher than the 0% ratio in Experiment 1) reported becoming aware of the temporal distortions during the experiment. The words they used to describe their subjective perceptions of the perturbations included "echo", "out of sync" and "garbled". However, there was no evidence that these six subjects' showed timing adjustment responses that were different from the other subjects.

The F2 trajectories produced by a representative subject (a 24 year-old male) under the noPert, Accel and Decel conditions are shown Fig. 2.15. The average noPert and Accel trajectories are shown in Panel B. From these average F2 trajectories, it can be seen that the Accel perturbation did not lead to substantial changes in the average timing of the F2 maxima and minima during the utterance. In contrast, as can be seen in Panels D and E of Fig. 2.15. the timing of these F2 extrema in the subject's production are substantially delayed under the Decel perturbation (darker green curve in Fig. 2.15.E) than the noPert baseline (black curve). This can be seen by comparing the F2 minima at $[u]_1$ and $[u]_2$, as well as by comparing the F2 maxima at $[j]_1$ and $[j]_2$. Therefore, the pattern of timing adjustment under the Decel perturbation was asymmetric. Whereas the Accel perturbation elicited little, if any, change in timing of the F2 landmarks, there seems to be a global rightward shift (i.e., delaying) of the F2 trajectories in response to the Decel perturbation.

On the group level, the subjects' articulation showed an asymmetric pattern of temporal changes under the Accel and Decel perturbations, as in the individual subject shown above. Significant articulatory timing changes were observed only under the Decel perturbation, which resulted in increases in both the [i]-$[u]_1$ and [i]-$[j]_1$ intervals. This can be seen from the slightly delayed F2 minimum at $[u]_1$ and F2 maximum at $[j]_1$ in the average Decel curve compared to those in the average noPert curve in Fig. 2.16.A. As Panels B and C of Fig. 2.16. show, the adjustments in the [i]-$[u]_1$ interval and the [i]-$[j]_1$ interval were quite small under the Accel perturbation, but were much greater and statistically significant under the Decel perturbation. The main effect of perturbation condition was significant for both intervals ([i]-$[u]_1$ interval: $F_{2,54}=13.38$, $p<0.0001$; [i]-$[j]_1$ interval: $F_{2,54}=16.38$, $p<0.00001$); the changes of both intervals under the Decel perturbation from the noPert baseline were statistically significant (Fig. 2.16.B and C).

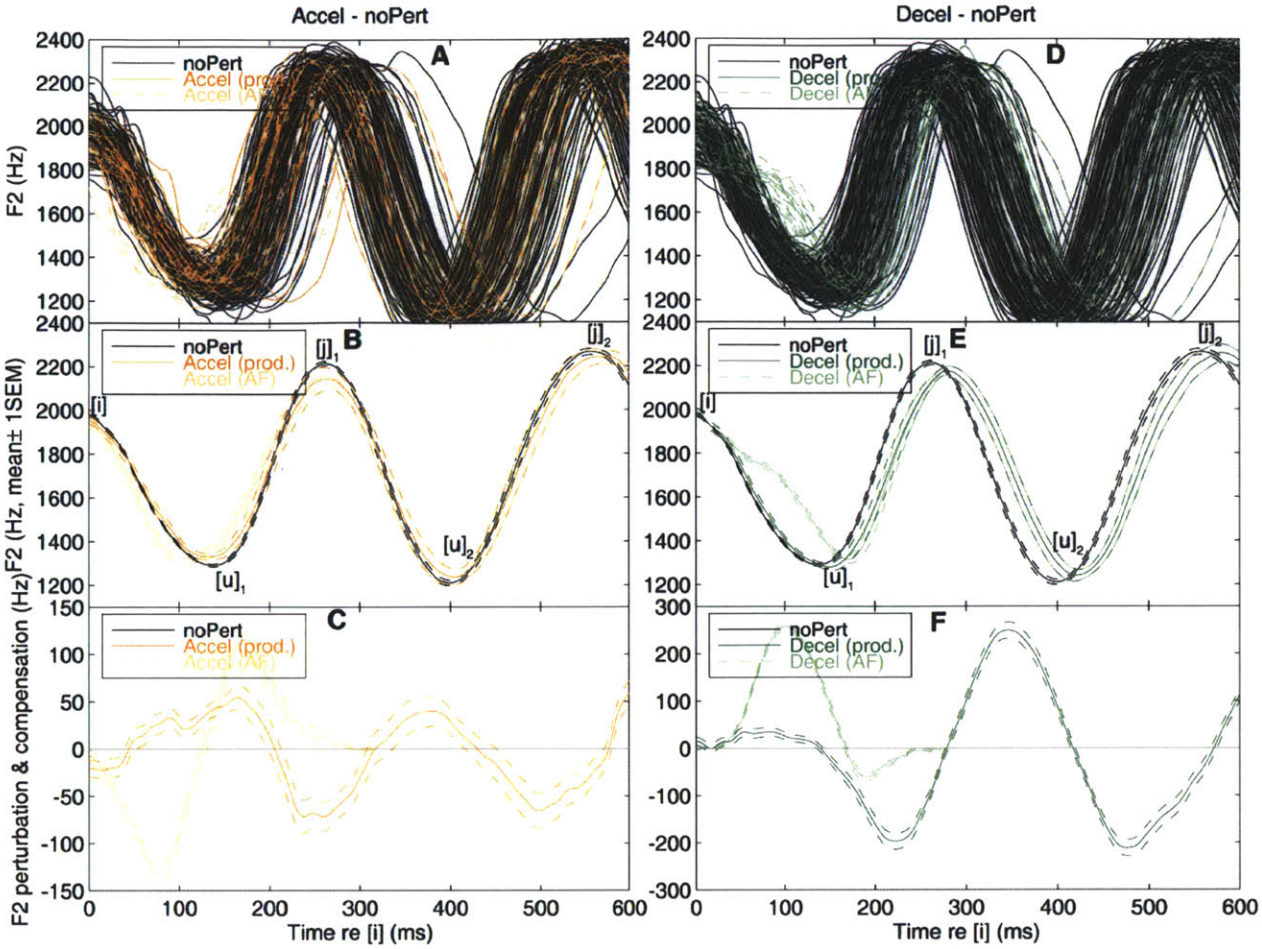


Figure 2.15. Compensatory changes in articulation in response to the Accel and Decel perturbations in a representative normal subject. **A.** F2 trajectories from individual trials. The two different conditions: noPert and Accel are shown in black and darker orange, respectively. The lighter orange line shows the perturbed F2 trajectories in the AF in the Accel trials. The trajectories are all aligned at the F2 maximum in [i]. **B.** Average F2 trajectories produced by the subject under the noPert (black) and Accel (orange) trials. The lighter orange curve shows the average F2 trajectory in the perturbed AF in the Down trials. **C**. Darker orange curve: the mean difference between the F2 trajectories produced under the noPert and Accel conditions. Lighter orange curve: the mean time course of F2 changes in AF caused by the perturbation. Note that the horizontal (time) axes of Panels A, B and C are identical. Panels **D**, **E** and **F** have the same format as Panels A, B and C, respectively, but show the Decel perturbation and the compensatory response to it by the same subject.

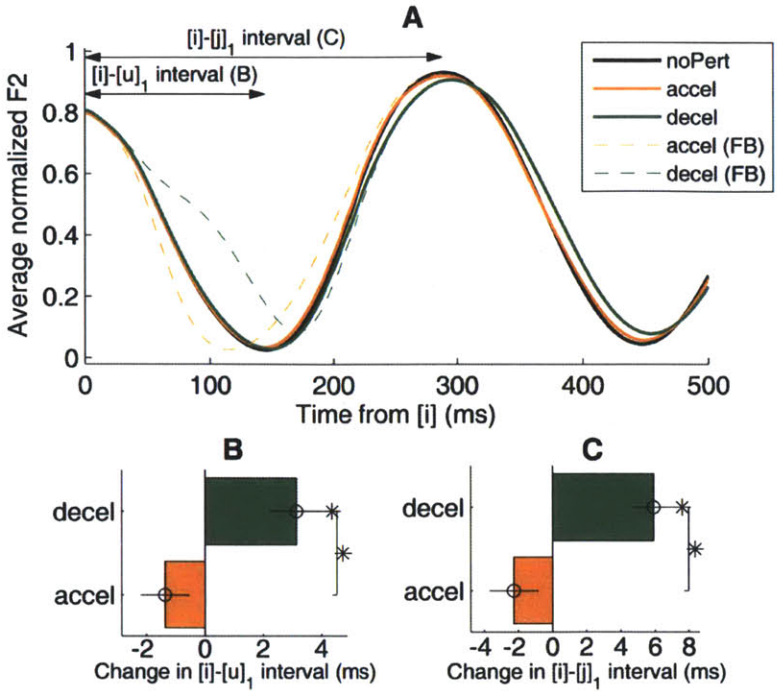


Figure 2.16. Articulatory compensations under the temporal (Accel and Decel) perturbations. **A**: grand average (across trials and subjects) of F2 trajectories aligned at the F2 maximum at [i]. The format is the same as Fig. 2.12A. The solid curves show production; the dashed curves show AF. The magnitude of the F2 at the $[u]_1$ minimum under the Decel perturbation (dashed green curve) appears to be altered by a substantial amount from the value in the production because the timing of the $[u]_1$ minimum varies across different trials and different subjects. In individual trials, the F2 magnitudes at this minimum were always preserved by the Decel perturbation (see Panel B). **B** and **C**: articulatory timing changes under the perturbations. D: change in the $[i]$-$[u]_1$ interval (error bars: +1 SEM). E: change in the interval between the $[i]$-$[j]_1$. Asterisks: significant difference at $p<0.05$ (post hoc Tukey's HSD following RM-ANOVA).

These temporal adjustments were qualitatively different from the spatial compensation observed in Experiment 1. The timing adjustments in this experiment *followed* the direction of the temporal change in the auditory feedback; whereas the spatial corrections in Experiment 1 *opposed* the feedback perturbations. Across the 28 subjects in Experiment 2, the ratio between the change in the $[i]$-$[u]_1$ interval in the subjects' production under the Decel perturbation and the perturbation of that interval in the auditory feedback was 14.7±4.0% (Mean ± 1 standard error of the mean). Similarly, the change in the $[i]$-$[j]_1$ produced interval amounted to 26.96±5.4% of the perturbation of the $[i]$-$[u]_1$ interval in the auditory feedback. These ratios of temporal

compensation are somewhat greater than the ratios of compensation under spatial perturbation observed in Experiment 1 and in previous studies that concentrated on static articulatory gestures (Purcell and Munhall 2006b; Tourville et al. 2008).

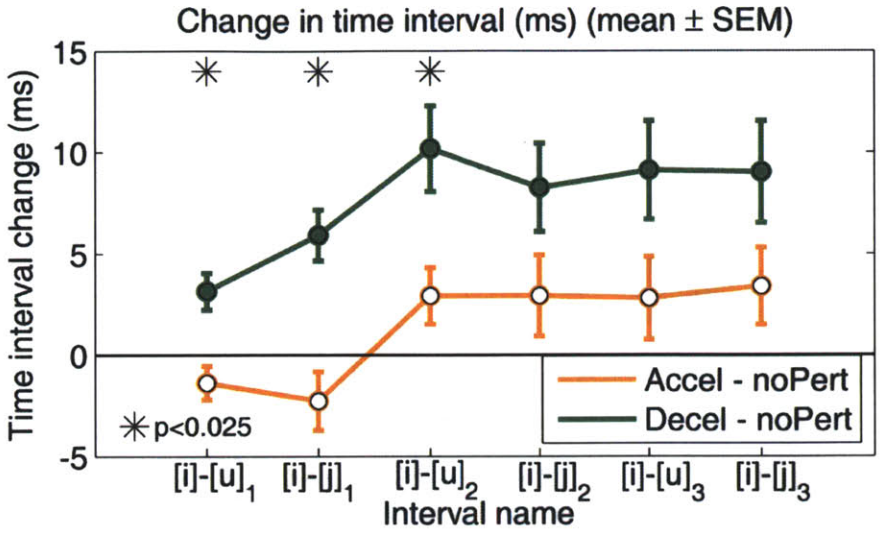


Figure 2.17. Changes in articulatory timing beyond the vicinity of the focus interval. Changes in the timing of the six major F2 landmarks ($[u]_1$, $[j]_1$, $[u]_2$, $[j]_2$, $[u]_3$ and $[j]_3$, see Table 1)under the Accel and Decel perturbations. The filled symbols represent significant difference from the baseline (one-sample t-test, $p<0.025$); the asterisks indicate significant difference between the Accel and Decel conditions (paired t-test, $p<0.025$).

In addition to the effects on the $[i]$-$[u]_1$ and $[i]$-$[j]_1$ intervals, which were relatively close in time to the perturbation interval, the Decel perturbation also caused timing alterations in later parts of the utterance. As Fig. 2.17. shows, the timing of the six major F2 landmarks (including the minima of $[u]_1$, $[u]_2$ and $[u]_3$, and the maxima of $[j]_1$, $[j]_2$ and $[j]_3$) all showed significant lengthening under the Decel perturbation. These results indicate that although the manipulation of AF was applied locally on an early part of the sentence, the Decel perturbation had global effects on syllable timing within this utterance. By contrast, the Accel perturbation caused no significant change in any of the three time intervals.

2. 3. Discussion

In this study, we performed two experiments which involved perturbations of speakers' AF of formant trajectories during the articulation of a multisyllabic. From measuring the trajectory of F2 produced by subjects under those different types of perturbations, we observed significant and specific acoustic adjustments (which reflect articulatory adjustments) in response to these perturbations. To our knowledge, this is the first study to provide evidence indicating that the speech motor system uses auditory feedback to fine-tune spatiotemporal parameters of multisyllabic articulation in an online, moment-by-moment basis during multisyllabic articulation and to characterize the spatiotemporal details of this online feedback-based control. *Experiment 1: responses to spatial perturbations.* The spatial perturbations used in Experiment 1 elicited significant changes in the magnitude of F2 in the production during and following the perturbation interval. These compensatory F2 adjustments are qualitatively similar to the previously observed compensation during the monophthongs [ɛ] (Purcell and Munhall 2006b; Tourville et al. 2008) and pitch compensation during word production and phonation (e.g., Burnett et al. 1998; Donath et al. 2002; Xu et al. 2004). However, since the compensatory responses in the current study were observed during time-varying articulation, they indicate that the role of auditory feedback in online articulatory control extends beyond the stabilization of static or quasi-static gestures, and to the control of articulatory trajectories that connect phonemes in a sequence.

The observed patterns of change are consistent with a control system that uses internal forward models (Miall and Wolpert 1996; Wolpert et al. 1998; Kawato 1999) to predict the future evolution of the acoustic parameters (F2 in the case of the current study) and preemptively correct the predicted errors. These forward models integrate sensory feedback (auditory feedback

in this case) with motor efference copies to predict the consequences of the motor programs that are about to be issued (e.g., Hickok et al. 2011). These predictions are compared with the auditory targets for the phonemes to be produced (Guenther et al. 1998). If a mismatch arises between the two (predicted errors and the auditory target), the control system will modify the motor programs before they are issued to the articulators, so as to preemptively minimize the errors. For example, under the Up perturbation, the subjects responded by lengthening the period ending in $[u]_1$ (i.e., the [i]-$[u]_1$ interval) relative to the noPert baseline (Fig. 2.12B). This changed can be explained by the artificially introduced upward F2 shift in the auditory feedback before the moment of $[u]_1$, which caused the forward models to predict a higher-than-needed F2 at $[u]_1$ with the unaltered motor programs. In response, the control system increased the duration of the gesture that led to downward sweep of F2, in order to counteract the predicted auditory error. Similar explanations apply to the $[u]_1$ F2 increase and the [i]-$[u]_1$ interval decrease under the Down perturbation, as well as to the $[j]_1$ F2 changes under both the Down and Up perturbations. In Chapter 3, these conceptual ideas will be implemented in a mathematically explicit computational model of speech motor control during multisyllabic articulation.

Tourville et al. (2008) observed that the bilateral posterior superior temporal cortex, right motor and premotor cortices, and inferior cerebellum are involved in the online auditory feedback-based control of a static articulatory gesture. We postulate that the online control of multisyllabic articulation involves a similar neural substrate, possibly with the additional role played by cerebellum in internal modeling and state estimation (Miall et al., 2007), which are necessary for forming sensory expectations during sequential movements.

To understand why the [i]-$[j]_1$ interval showed smaller changes than the [i]-$[u]_1$ interval, we may consider the reversal of the direction of F2 at $[u]_1$. In the case of the Up perturbation, this

reversal causes the *undershoot* predicted at $[u]_1$ to lead to the prediction of an *overshoot* at the next extremum, $[j]_1$. Hence the temporal corrections that might have been made during the periods before and after the reversal may have canceled each other, leading to the non-significant change observed in the [i]-$[j]_1$ interval.

The value of the compensatory adjustment in the magnitude of the produced F2 was approximately 14% of the magnitude of the perturbation in the auditory feedback (Fig. 2.10). This ratio of compensation appears to be larger than the ratio of compensation observed in the prior studies of the monophthong [ɛ], which was shown to be around 3-6% at 250 ms after perturbation onset by Tourville et al. (2008) and Purcell and Munhall (2006b). This result may reflect a greater role of AF during time-varying articulation and phoneme-to-phoneme transitions than during within-phoneme articulatory gestures, and appear to be consistent with the finding of Xu et al. (2004) and Chen et al. (2007), who observed greater compensations to perturbations of pitch feedback in dynamic tonal sequences than in static (repeating) ones. Therefore, there seems to be converging evidence for a greater role of AF-based control during the production of sequential or time-varying gestures than during the task of prolonging quasi-static articulatory or phonatory gestures.

Experiment 2: response to temporal perturbations: The temporal compensation observed under the Accel and Decel perturbations of Experiment 2 altered the timing of the local F2 minimum that corresponds to $[u]_1$ in the word "owe" in the auditory feedback. One of the two types of perturbation, Decel, led to not only a significant lengthening of the syllable [ou] ("owe") in the subjects' production, but also delayed initiation of the following syllable [ju] ("you"). These temporal corrections were small in absolute magnitude, but accounted for considerable fractions (15-27%) of the timing perturbations in the auditory feedback. In addition, the timing of the

syllables subsequent to the cessation of the Decel perturbation was also altered. These findings argue against the notion that the syllable onset timing in an utterance is completely pre-programmed and determined by processes unrelated to auditory feedback (Fowler 1980), which is adopted by the Task Dynamic model of speech articulation (e.g., Saltzman and Munhall 1989; Saltzman et al. 2006). Contradictory to this concept of a "timing score" that completely determines the timing of syllables, not unlike the role of a musical score in specifying the timing of musical notes, our findings provide further evidence against completely pre-programmed timing and support the notion that articulatory timing can be adjusted dynamically as the sensorimotor process of articulation unfolds. In particular, the speech motor system may process the auditory feedback from earlier segments or syllables of an utterance in a way to generate information that is used for some of the guidance of the articulatory timing in ensuing parts of speech. Hence any model of the neural mechanisms of the control of multisyllabic articulation need to incorporate AF-guided online fine-tuning of syllabic timing. The sqDIVA model which we will develop in Chapter 3 of this thesis will be a model that meets this requirement.

The response to these temporal perturbations showed an asymmetric pattern: whereas the Decel perturbation led to significant delays in the termination of the perturbed syllable and the initiation of the following syllables, the Accel perturbation elicited little, if any change in articulatory timing in the subjects' production. This asymmetric pattern of timing adjustment is consistent with the previous observation by Perkell et al. (2007) that whereas sudden loss of AF (by switching off the cochlear implants worn by the subjects) during production of a vowel led to significant lengthening of the duration of the vowel, sudden restoration of auditory feedback (by switching on the cochlear implants) caused no significant changes in vowel duration. More recently, Mochida et al. (2010) also observed asymmetric temporal compensation to temporal

perturbations in a group of subjects who repeatedly produced the nonsense syllable [pa], but their results differed slightly from the findings of the current study. They observed significantly earlier-than-baseline initiation of syllables in response to auditory feedback advanced in time, but no significant change in production timing under a delayed auditory feedback. These apparently contradictory findings may be attributable to the different nature of the speech tasks. The current study used linguistically meaningful utterances with varied syllables and only vocalic phonemes, as well as a self-generated, close-to-natural speaking rate. By contrast, Mochida and colleagues used externally auditory clicks to pace the rhythm of the production and used nonsense utterances consisting of repeating syllables and voiceless plosives. Future studies are needed to examine how linguistic, phonological and prosodic factors may affect the interaction between auditory feedback and articulatory motor control. The neural substrates of the feedback-based timing adjustments may include the basal ganglia and cerebellum, which both have been shown to play roles in speech motor timing (e.g., Wildgruber et al., 2001; Ackermann, 2008).

The effects of noise masking (Lane and Tranel 1971; Van Summers 1998) and delayed auditory feedback (DAF) (Fairbanks 1955; Zimmermann et al. 1988) on temporal parameters of connected speech have long been known. Both manipulations lead to slowing down of speaking rate; furthermore, DAF can lead to breakdowns of speech fluency. However, the interpretation of those results has been controversial. Arguments against interpreting those data as supporting a role of auditory feedback in multisyllabic articulation have been based mainly on the sustained nature and unnaturalness of the noise-masking and DAF conditions, which may "force" the speaker to attend to the auditory feedback and not necessarily reflect the control strategy used under normal (unperturbed) speaking conditions (Lane and Tranel 1971; Borden 1979). The

perturbations used in the current study were subliminal in comparison with the readily perceived traditional manipulations of auditory feedback. Most subjects in the current study reported being unaware of any deviations of auditory feedback from the normal pattern. Therefore it seems reasonable to assume that the patterns of compensation observed under the perturbations imposed in this study can be more readily interpreted as reflecting mechanisms used in unperturbed speech production.

The current findings demonstrate that the normal process of speech motor control makes use of auditory feedback to optimize the articulatory process, with the aim of minimizing the amount of error in reaching the auditory goal regions (Guenther et al. 1998; Guenther 2006; Matthies et al. 2008) for successive phonemes and to achieve the intended temporal relationships between the phonemes. This view is compatible with the deterioration of the acoustic precision of produced speech that is observed when auditory feedback is unavailable to the speaker, as when high-intensity noise masking leads to decreases in the phonemic contrasts between vowels and between sibilant consonants (Van Summers et al. 1988; Perkell et al. 2007) and when auditory masking increases the variability of the relative timing between articulatory events in speech (Namasivayam et al. 2009).

Previous studies based on unanticipated mechanical perturbation of the lips and jaw during speech demonstrated that the speech motor system exhibit short-latency, task-specific compensations to mechanical perturbations of the articulators (e.g., Abbs and Gracco 1983; Gracco and Abbs 1985; Munhall et al. 1994; Shaiman and Gracco 2002). Similar to the results of the current study, the mechanical perturbations can cause compensations in both the magnitude and timing of the articulatory movements (Gracco and Abbs 1989). When viewed in light of those previous results, the results of the current study seem to indicate that the speech motor

system makes use of both somatosensory and auditory feedback to control articulatory movements online. Honda and colleagues (2002) examined the interactions between these two modalities of sensory feedback in speech motor control by introducing perturbations of palate shape (by an inflating a small balloon attached to a palatal prosthesis) during a subject's productions of sequences of the syllable /ʃa/. In response, the subject made compensatory adjustments of the tongue position for the fricative /ʃ/ that were 25-50% smaller when auditory feedback was masked by noise than without masking. This finding is consistent with an online control system of articulation in which auditory and somatosensory feedback both make contributions.

Existing models of speech motor control vary in their ability to predict the online compensatory responses observed in the current study. The Task Dynamic model (Saltzman 1989; Saltzman et al. 2006) is capable of specifying spatiotemporal details of speech motor events by virtue of its use of pre-determined gestural scores. However, this control mechanism in this model is entirely feedforward. So in its current form, the Task-dynamic model has no capability for simulating the online, feedback-based adjustments of articulatory magnitudes and timing seen in our data. The DIVA model (Guenther 2006) proposed a control scheme incorporates both feedforward and feedback-based control mechanisms and has been shown to successfully predict the online F1 compensation under auditory perturbation during the production of the monophthong [ɛ] (Tourville et al. 2008). However, this model in its current form is restricted to control the production of simple, short utterances, i.e., *articulatory units* such as single syllables and frequently used short words, and lacks the capability to model the spatiotemporal patterns of multisyllabic, connected speech. The state feedback control (SFC) model (Ventura et al. 2009; Hickok et al. 2011) has the potential to explain the current data set,

but such models are still in the form of block diagrams and hence lack the computational detail and predictive power of models such as DIVA. Future work is needed to develop the mathematical details of a state-feedback control model. Such work will contribute to a deeper understanding of speech motor control and motor control in general. In Chapter 3, we will develop a new model that will fill the gap of modeling work related to the interaction of AF and movement control in multisyllabic speech articulation called sqDIVA.

The online adjustments of the articulatory positions and timing that can be inferred from our data are consistent with Levelt's (Levelt 1989) notion that speakers attend to and monitor virtually every aspect of the speech production process. In the current chapter, we discovered a type of monitoring that has not been described before: the speech motor system monitors the spatiotemporal details of auditory feedback, extracts relevant information from them during rapid sequencing of phonemes and syllables, and then use such information to fine-tune both the spatial and temporal parameters of the ensuing speech movements with a short latency.

These findings raise the question of whether state-feedback control may be operating during the production of other types of highly skilled sequential movement such as musical performance, cursive handwriting and keyboarding, which are also likely to be influenced by multisensory feedback. As in the case of speech production, devising paradigms to investigate this issue offers interesting theoretical and experimental challenges.

Chapter 3. Computational modeling of auditory-motor interaction in multisyllabic articulation

To our knowledge, the experimental data presented in Chapter 2 are the first evidence to support a role of auditory feedback in the online control of spatiotemporal parameters of time-varying articulatory trajectories. The Down-Up and Accel-Decel perturbations used to manipulate the auditory feedback led to not only changes in the values of F2 along the trajectories in the subjects' production, which reflect the compensatory adjustments of the articulatory positions, but also adjustments in the times at which these phoneme-related positions were attained in the subjects' articulation. Therefore it is tempting to conclude that auditory feedback plays roles in the online control of both the spatial and temporal parameters of articulation. However, since the spatial and temporal aspects of a time-domain trajectory are intricately intertwined, it is possible that the significant but small temporal corrections observed under the perturbations were merely by-products of certain control processes not directly related to timing or the sequencing of multiple articulatory units (e.g., syllables). For example, Tourville et al. (2008) generated quantitatively accurate fitting of the online F1 compensation patterns during the monopthong [ɛ] with the DIVA model, which in its current form, has no components or mechanisms explicitly related to sequencing or the timing of multiple syllables. It is possible that the complex compensatory patterns we observed in Chapter 2 can be explained adequately by this type of simple models that do not deal directly with timing control.

Without quantitative modeling the auditory feedback-based articulatory control, it will be hard to prove or disprove this possibility, because formant trajectory in our multisyllabic utterance are much more complex than the quasi-constant formant trajectories used in previous monophthongs perturbation studies (Purcell and Munhall 2006b; Tourville et al. 2007). Despite the fact that the stimulus utterance we chose ("I owe you a yo-yo") was kept intentionally simple in its phonetic composition, the F2 trajectory of the stimulus utterance, with its multiple inflections, is much more complex than the essentially constant F1 value in monophthongs used

in previous studies. As a consequence, the patterns of compensatory F2 changes are considerably more complex than previously seen. This can be appreciated by looking at Figures 2.6., 2.7. and 2.15. Given this level of complexity, computational modeling is a powerful and possibly the only way through which we can proceed beyond the level phenomenological data and attain insight about the detailed organization principles of online multisyllabic articulatory control based on AF.

3.1. Existing models of sensorimotor articulatory control

There is a long tradition of modeling feedback control in the speech motor system (e.g., Fairbanks 1954; Mysak, 1960; Neilson and Neilson, 1987). However, due to the prior lack of mathematical details, it is generally impossible to make concrete, quantitative predictions about responses to specific auditory feedback perturbations using such models. Therefore, these early models remained vague and difficult to test. To our knowledge, the only mathematically defined model that deals explicitly with the sensory-motor interaction in the speech motor system is the DIVA model (Guenther, 2006; Golfinopoulos et al., 2009). According the DIVA model, two modes of motor control, namely feedforward and feedback control function in parallel during the production of speech sounds. The feedforward pathway reads out previously learned motor programs and issues commands to the articulators to direct the velocity (speed and direction) of the articulatory movements. As a part of the feedback pathway, the auditory system monitors the acoustic consequences of the articulation and compared the auditory feedback with the auditory target for the sound being produced. Mismatches between the auditory feedback and the auditory target give rise to auditory errors. These errors are processed by the feedback control map, hypothesized to reside in the right ventral premotor cortex (Tourville et al., 2008, Golfinopoulos et al., 2009), to generate corrective motor commands that can nudge the production in a direction opposite to the auditory error and hence compensate for the error. With proper tuning of a small number of parameters, the DIVA model is capable of generating quantitatively accurate

predictions about the online compensation to perturbations of auditory feedback during the production of the monophthong [ɛ] (Tourville et al., 2008).

However, an important limitation of the current DIVA models is that it has been designed to deal primarily with articulation during single "units" of speech, such as single syllables (e.g., /ba/) and frequently used short words and utterances (cf. the "mental syllabary" of Levelt and Wheeldon, 1994). DIVA, in its current form, is poorly posed to answer questions related to multisyllabic articulation. Although DIVA has been used to model short utterances comprised of multiple syllables, such as "good doggie" (Guenther, 2006), the timing in the production of these multisyllabic utterances by the DIVA model is implemented in a way that treats the multisyllabic utterance such as "good doggie" as a unitary entity, i.e., a long and complex "syllable", instead of as a sequence of discrete syllables, e.g., /gud/, /dɑ/, and /gɪ/. Through practice, DIVA can produce this utterance intelligibly (Guenther, 2006), but it has not been determined whether DIVA is able accurately predict the compensatory responses in articulation if multisyllabic utterances are produced under perturbation of sensory feedback. This is an important question because in essence, it underlies and motivates a test of the validity of the current DIVA model in modeling multisyllabic articulation.

The experimental data from the current study provides an opportunity for carrying out such a test. If the utterance "I owe you a yo-yo" is indeed regarded by the speech motor system as a single unit for control, the current DIVA model, if correct about the sensorimotor processes during connected speech, should be capable of accurately predicting the experimentally observed compensatory patterns, including the F2 value changes and the timing changes. If this is the case, we will have no reason to question the control scheme in the current DIVA model. However, if the DIVA model fails to accurately predict the spatial and temporal compensations observed in the perturbation experiment, the way the current DIVA model deals with multisyllabic utterances will be falsified, and new, alternative models should be sought to model the experimental data.

3.2. The sqDIVA model

The way DIVA currently deals with multisyllabic utterances suffers from a conceptual weakness: the vast (virtually infinite) number of possible utterances in a language leads to a combinatorial explosion (cf., Norman, 1980) of the motor programs to learn and store. It is highly unlikely that the speech motor system stores pre-learned motor trajectories for all possible utterances. Many of the utterances we produce in daily life are ones that we have never produced before. How is it possible that we can produce those utterances that are new to our speech motor system even though we haven't practiced or even heard them before? Apparently, it is a much more parsimonious and plausible approach to 1) store the auditory targets for single syllables (or other types of units), and 2) use a sequencing mechanism to string the units together during the production of multisyllabic utterances.

Here we propose a simple sequencing mechanism for the production of multisyllabic utterances. We dub this alternative model sqDIVA ("sq" is an abbreviation for "sequential"). This control mechanism in sqDIVA is based on a premise that is similar to the one DIVA is based on, namely that the primary goals of articulation are in the acoustic/auditory domain (at least for vowels, Perkell et al., 1997, Guenther et al., 1998) and that the speech system monitors auditory feedback online for correcting articulatory errors.

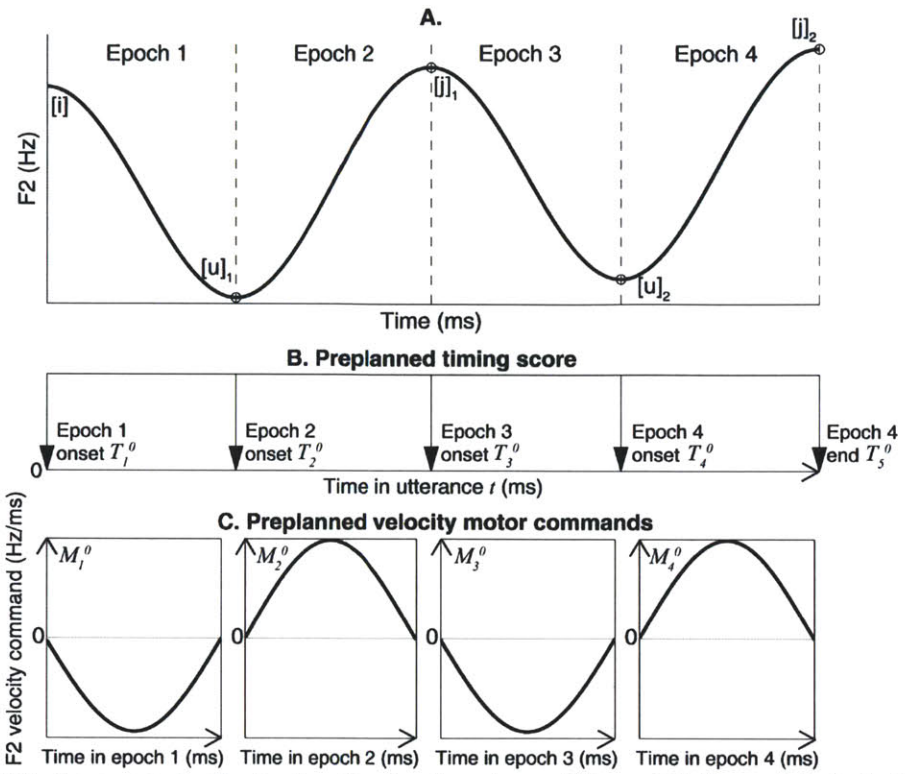


Figure 3.1. An example illustrating the basic set-up of the sqDIVA model. **A:** In this example, the target F2 trajectories for the four epochs are shown by the solid curve. **B:** An example showing the preplanned timing score. **C:** the preplanned velocity motor commands. See text for details of the model.

Figure 3.1.A uses a part of the utterance "I owe you a yo-yo" as an example for showing the sequencing mechanism in sqDIVA. The time axis is divided into a number of so-called *epochs*. An epoch is defined by a period in which there is a monotonic transition in the formant value. Some of the epochs contain a syllable and others are the transitional periods between two adjacent syllables. For example, in Fig. 3.1.A, part of the first epoch contains the syllable "owe", which involves a monotonic decrease in the value of F2; the second epoch corresponds to the transition from the end of the syllable "owe" to the beginning of the following syllable "you", which involves a monotonic increase in F2; and so on. We hypothesize that before initiation of the utterance, the onset timing of all the epochs are preplanned (Fig. 3.1.B). This timing preplanning hypothesis is similar to the preprogrammed timing in the Task Dynamics (TD) model (Saltzman et al. 2006, Namasivayam et al. 2009; see also Fowler 1980). Preplanned timing is a simple but useful hypothesis which is necessary for explaining the fact that the

speaking rate and syllable timing can be controlled by a speaker voluntarily with a high degree of flexibility during speech (c.f., prosodic control in speech, stress patterns, chanting and singing). However, as we will see below, the timing plan in our models is not a rigid, unchangeable one (Keele 1968), but instead incorporates sensory feedback in the online control of articulation.

It should be noted that the in this model, the end of the onset for each epoch is the onset of the following one. This scheme of modeling event timing was devised for the sake of simplicity. It is clear from previous empirical observations and theories on the phenomenon of coarticulation (Öhman 1966, 1967, Daniloff and Hammarberg 1973, Fowler 1980, Keating 1988, Farnetani and Recasens 1999) that there are interactions among and possible overlapping of different articulatory gestures in time. Future extensions of our model will incorporate this overlapping of timing relations between consecutive articulatory events, but the underlying principles of online spatiotemporal control should not be affected in a fundamental way by such changes.

In addition to the preplanned epoch onset timing, the model prepares a set of articulatory commands for each epoch. For clarity and computational tractability, we opted to ignore the geometrical details of the vocal tract and the complexities of the relations between vocal-tract configuration and acoustic parameters (e.g., formants) of the produced sound, while retaining the capacity to show the organizational principle of online spatiotemporal adjustments. Specifically, this model directly controls the evolution of F2 in time by using "motor commands" expressed in terms of F2 velocity. Examples of these epochal velocity commands are shown in Fig. 3.1.C. This approach follows DIVA's use of velocity control. The velocity commands for the within-syllable epochs (e.g., epochs 1 and 3 in Fig. 3.1.) can be regarded as a set of feedforward motor commands acquired through feedback-driven learning, similar to the learning in DIVA (Guenther 2006). The velocity motor commands for the between-syllable epochs may be regarded as either pre-learned or as computed "on the fly" through inverse modeling (Kawato

1999). Since the between-syllable epochs are defined uniquely by the phoneme at the end of the preceding epoch and the phoneme at the beginning of the following one, the number of such transitional epochs should be on the same order of magnitude as the number of phonotactically legal syllables in a language, and hence do not suffer from the combinatorial explosion problem as mentioned above.

Given the above-described setup of the model, its mathematical formulation is as follows. We denote the value of F2 at time t as $f(t)$. The initial value of $f(t)$ at time 0 (onset of the articulation) is determined by the following equation:

$$f(0) = \hat{F}_1^0(0), \qquad (3.1)$$

which sets the initial condition of the model. $\hat{F}_1^0(0)$ is the target formant value at the onset of the first epoch, which is known a priori by the model, given the phonemic content of the utterance. The subscript 1 represents the first epoch, and the superscript 0 indicates that it is the version of $\hat{F}_1$ at time $t = 0$ that is being used. The purpose for including a superscript here is to allow the online updating of the targets and commands, which will be described below.

Suppose the i-th epoch is the currently produced epoch, the first derivative (velocity) of the formant $\dot{f}(t)$, is:

$$\dot{f}(t) = M_i^t(t - T_i^t) + M_{FB}(t), \text{ for } T_i^t \leq t < T_{i+1}^t, \qquad (3.2)$$

In Equation 3.2, M_i^t is the velocity command for the i-th epoch; the superscript t indicates the latest version of the command, i.e., the version at the current time t. T_i^t is the onset time of the i-th epoch and T_{i+1}^t is the beginning of the next $(i + 1)$-th epoch and the end of the current i-th. Similar to before, the superscript t indicates the latest value of this onset timing. These t's in the

superscripts are necessary because as we will show later, both the velocity command M and timing score T are updated online by using information form AF. $M_{FB}(t)$ is the feedback command and will be described below.

The model constantly monitors the produced F2 in the AF. At time t, the latest AF value of F2 the model can act on is from time $t - D$, denoted as $a(t - D)$, where D is the "round-trip" latency of the auditory feedback pathway of the speech motor system, a free parameter in the sqDIVA model. Prior formant- and pitch-perturbation studies have provided a range of this latency between 80 and 200 ms (e.g., Burnett and Larson, 2002, Donath et al., 2002, Natke et al., 2003, Tourville et al., 2008, Chen et al., 2010). Under a normal and unperturbed condition, the feedback equals the actual production:

$$a(t - D) = f(t - D), \quad (3.3)$$

However, artificially introduced manipulation of AF, as the ones used in this study, can alter $a()$.

The model maintains and updates a target trajectory for the epoch being produced, denoted as $\hat{F}_i^t(t)$, which we have seen in Equation 3.1 before. This target trajectory is similar to the auditory targets in DIVA, which are learned through the auditory perceptual channel during speech acquisition. As the sqDIVA model is primarily concerned with the online feedback-based control of articulation, this model will not account explicitly for processes of learning these target trajectories. As for the motor command and the timing score, the superscript t denotes the latest version of the target. The target formant value is compared with the value in the auditory feedback to give rise to the estimated feedback error:

$$e(t - D) = a(t - D) - \hat{F}_i^t(t - D - T_i^t), \quad (3.4)$$

Note that T_i^t is the onset time of the current (i-th) epoch.

An important feature of the sqDIVA model is the presence of two parallel feedback pathways which both utilize this error signal to perform online correction of articulation. We refer to these two feedback-based control schemes as the *simple* and *complex* corrections. The simple correction is similar to the feedback control scheme from the DIVA model and is the one that is responsible for generating the feedback velocity command M_{FB} in Equation 3.2. M_{FB} is simply the feedback error scaled by a given factor w_{FB} :

$$M_{FB}(t) = -w_{FB} \cdot e(t - D), \qquad (3.5)$$

w_{FB} is a second free parameter of the model and it specifies the strength of the simple correction.

The model described so far includes only two free parameters (D and w_{FB}). It can be regarded as a simplified version of the DIVA model. This two-parameter model doesn't yet include any component for the online correction of timing. We will use this model as the ***baseline*** model, which will be compared with more complex variations of the model.

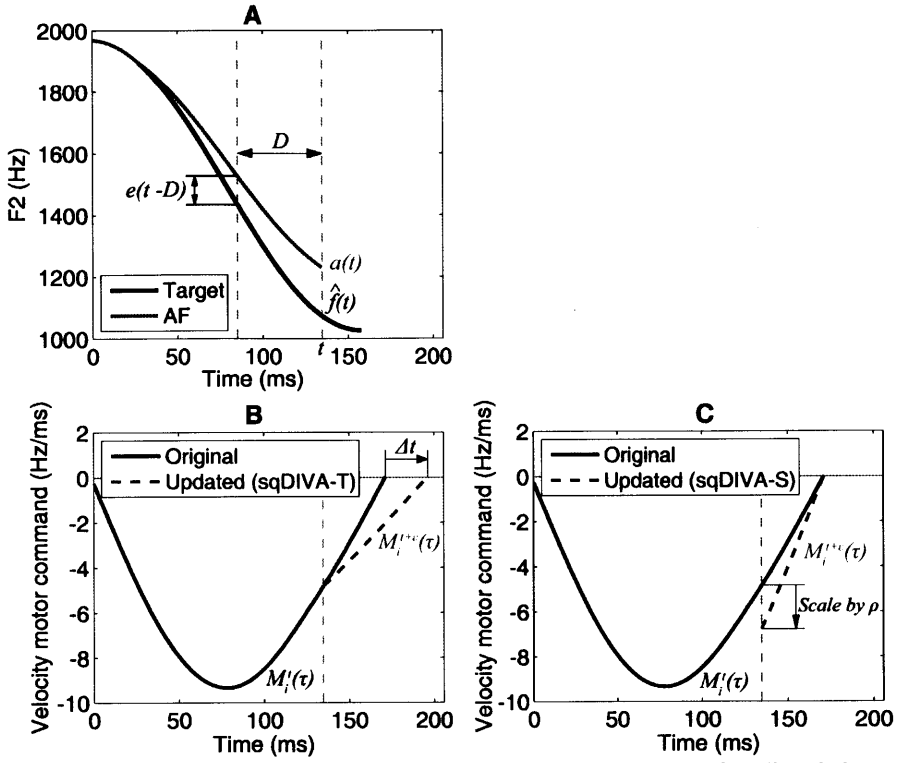


Figure 3.2. A schematic example illustrating the online auditory feedback-based correction of temporal or spatial aspects of articulation. **A:** The auditory feedback shows an undershoot relative to the target. Panels B and C show two possible schemes of correcting for this error. **B:** the temporal correction in sqDIVA-T, in which the onset time of the next epoch is shifted to the right (i.e., delayed) and the velocity command stretched out in time to compensate for the undershooting error**C:** the spatial correction in sqDIVA-S, in which the motor command is scaled up in magnitude in response to the error.

The complex correction in this model involves shifting the epochal onset times to make them occur earlier or later according to the nature of the feedback error. Whereas the simple correction reacts to past errors, the complex correction acts proactively to minimize future errors. If the feedback error involves an undershoot with respect to the direction of F2 of the current epoch, the duration of the current epoch will be lengthened and the velocity command profile will be "stretched out" in time, so that the model will take more time to finish the articulation of the current epoch, which will, in turn, preemptively minimize the extent of the projected undershoot. Fig. 3.2.B schematically shows an example of this timing shift and "stretching-out" of motor command. It should be noted that the amount of timing adjustment in Fig. 3.2.B is exaggerated for visualization. As the duration of the current epoch is lengthened by a certain amount Δt, the onsets of all following epochs will be delayed by Δt. Vice versa, if feedback error involves an

overshoot (not shown), the model will shorten the duration of the current epoch and "compress" and velocity command in time, so that the projected overshoot can be minimized. Formally, the amount of temporal correction is determined by:

$$\Delta t = \begin{cases} +r_{BW} \cdot \delta(i) \cdot w_T \cdot \frac{e(t-D)}{\bar{M}_i}, if\ e(t-D) \cdot \bar{M}_i \leq 0 \\ -r_{BW} \cdot \delta(i) \cdot w_T \cdot \frac{e(t-D)}{\bar{M}_i}, if\ e(t-D) \cdot \bar{M}_i > 0 \end{cases}, \quad (3.6)$$

In the above equation, $\bar{M}_i$ is the average formant velocity in the i-th (current) epoch; $\delta(i)$ is a 0-1 indicator function which indicates whether the i-th is primarily within- or between-syllable. It takes the values of 0 and 1 for within- and between-syllable epochs, respectively. There are two free parameters in Equation 3.6. w_T is the temporal adjustment coefficient, which specifies the magnitude of the temporal adjustment. r_{BW} is the between-/within-syllable ratio, which is incorporated to accommodate the possibility that this temporal adjustment may be implemented with less strength during the epochs that are primarily between-syllable transitions (e.g., epochs 2 and 4 in Fig. 3.1.A) and than those that are contained within syllables (eg., epochs 1 and 3 in Fig. 3.1.A). This possibility may arise as a consequence of the assumption that it is the syllables, not the transitions between them that carry the linguistic information and hence require more careful online articulatory control. In this model, r_{BW} takes a real-number value between 0 and 1. At the two extremes, $r_{BW} = 0$ corresponds to a case in which no temporal adjustment is made during the between-syllable epochs, and $r_{BW} = 1$ corresponds to a case in which the temporal adjustment is made with the same magnitude for the two-types of epochs. When this timing correction occurs, the target trajectory of the current epoch $\hat{F}_i^t$ will be also stretched or compressed with by equal amount.

The sqDIVA-S model: an alternative for comparison

With this timing adjustment mechanism incorporated, the full sqDIVA model has four free parameters. The two additional free parameters that the Baseline model doesn't have are the

temporal adjustment coefficient w_T, and the between-/within-syllable ratio r_{BW}. Because this model involves online timing correction, we will refer to it as ***sqDIVA-T***. When comparing the performance of the sqDIVA-T model with the Baseline model, it may be argued that any gain in the simulation performance may be trivially due to increased degree of freedom (DOF) (i.e., increased number of free parameters). For this reason, we will introduce an alternative variant of the sqDIVA model, which we will refer to as the ***sqDIVA-S*** model. The sqDIVA-S model has the same DOFs as sqDIVA-T but performs the online correction in a different way. The "S" in the name of the model stands for "spatial" and its meaning will be clear from the description below. The performance of sqDIVA-S and sqDIVA-T, which both have the same number of DOFs, will be compared.

In the sqDIVA-S model, the complex correction scales the magnitude of the velocity commands, rather than shifting the epoch boundaries along the time axis (as in sqDIVA-T). If an undershoot is detected in AF, the velocity commands will be scaled up (i.e., increased in magnitude); vice versa, a scaling-down will be implemented in response to an overshoot in AF. The ratio of the scaling ρ is determined by:

$$\rho = \begin{cases} 1 + r_{BW} \cdot \delta(i) \cdot w_S \cdot \frac{e(t-D)}{T_{i+1}^t - T_i^t}, if\ e(t-D) \cdot \bar{M}_i \leq 0 \\ 1 - r_{BW} \cdot \delta(i) \cdot w_S \cdot \frac{e(t-D)}{T_{i+1}^t - T_i^t}, if\ e(t-D) \cdot \bar{M}_i > 0 \end{cases}, \quad (3.7)$$

in which $T_{i+1}^t - T_i^t$ is the duration of the current (i-th) epoch. The model parameter w_S is the spatial correction coefficient, which is analogous to the temporal correction coefficient w_T but specifies the magnitude of the feedback-based scaling. As in sqDIVA-T, we include the between-/within-syllable ratio r_{BW} to model the differential online adjustment during two different types of epochs. This scaling factor ρ is then used to modify the velocity command

$$M_i^{t+\varepsilon}(\tau) = \rho \cdot M_i^{t+\varepsilon}(\tau),$$

wherein ε is an infinitesimal time increment. An example of this spatial scaling is provided in Figure 3.2.C, which illustrates an up-scaling of the velocity commands in response to an undershoot in auditory feedback.

The three models (baseline, sqDIVA-S and sqDIVA-T) were fit separately to the experimental data. For each model, parametric space searches were made separately for the 30 participants of the behavioral study. The average F2 trajectory in a subject's production under the noPert condition was used to generate the target F2 trajectory ($\hat{F}_i$), the preplanned timing score (T_i) and the velocity motor commands (M_i). The same type of F2 feedback perturbation as used in the psychophysical experiments was applied to the model. The model fitting criterion was to minimize the cumulative difference between the model-simulated F2 compensation profile and the experimentally measured F2 compensation profile for the particular subject. In other words, a unique set of free parameters was fitted to every single subject. This approach was chosen because it is well known that different speakers respond in substantially individualized ways to the same manipulation of auditory feedback (e.g., Villacorta et al., 2007, Munhall et al., 2009), which may reflect individual difference in the properties of the speech motor and perceptual system.

A time step of 0.2 ms was used in the numerical time-domain simulation. To search for parametric optima, we used a large-scale interior-point algorithm (Waltz et al., 2006), implemented as the *fmincon* command for in the MATLAB Optimization Toolbox. In order to ensure the capturing of the global optimum, 1024 points, uniformly distributed in the following ranges of the parametric space, were used as initial conditions of the optimization procedure on sqDIVA-T:

$$D\epsilon[80,200]ms;$$

$$w_{FB}\epsilon[0,6\times10^{-4}];$$

$$w_T\epsilon[0,1.2\times10^{-2}];$$

$$r_{WB}\epsilon[0,1].$$

These ranges were selected empirically to ensure a complete coverage of the reasonably possible parametric values. For the sqDIVA-S model, the covered ranges of D, w_{FB} and r_{WB} are similar to above, and the ranges of the spatial correction coefficient is $w_S\epsilon[0,5\times10^{-5}]$. We chose these ranges of parameters through careful heuristic trials to ensure the coverage of the global optimum.

The following is a summary of the important features and properties of the two variants of the sqDIVA model.

1) To ensure continuity of theoretical modeling work, the sqDIVA model follows the basic principles of DIVA, including a) the use of velocity as the controlled variable and b) segregation and collaboration of the feedforward and feedback pathways.
2) However, unlike the DIVA model, sqDIVA model treats a multisyllabic utterance as a sequence of articulatory units, of which the onset and offset timing are explicitly controlled by the model.
3) The sqDIVA model incorporates the new functionality of online sequence adjustment. The two variants of sqDIVA have different ways of adjusting the sequence plan: the sqDIVA-T model adjusts the timing of the future syllable onsets and offsets by using information related to the overshooting or undershooting in the AF, whereas the sqDIVA-

S model adjusts the magnitude of the ensuing velocity commands without explicitly altering timing.

4) Both variants of sqDIVA have as few free parameters as possible, to conform to the principle of Occam's razor. To this end, simplifications are adopted. For example, details of the articulatory-acoustic relations and coarticulation between syllables are ignored. Also, as a first step, learning of the within syllable trajectories and the timing pattern is not incorporated in the first stage of the modeling.

3.3. Modeling Results

3.3.1. Modeling of the Up and Down perturbations

Both the sqDIVA-T and sqDIVA-S models have two more free parameters than the Baseline model. In the sqDIVA-T model, the two additional free parameters play the role of online timing adjustments, whereas in the sqDIVA-S model, the extra DOFs are used for online updating of command magnitude (i.e., spatial corrections). Not surprisingly, both the sqDIVA-T and sqDIVA-S model generated significantly better fitting of the individual subjects' compensatory trajectories than the Baseline model (paired t-test: $p=0.00002$ for sqDIVA-T and $p=0.00006$ for sqDIVA-S; see Fig. 3.3.A and 3.3.B). In Fig. 5A, the diagonal line indicates equal fitting errors of the sqDIVA-T and Baseline models; the data points above the diagonal line correspond to cases in which the sqDIVA-T model fit the experimental data better than the Baseline model. A similar format is used in Fig. 3.3.B. On average, the performance gains (i.e., decreases in the fitting error) were 23.3% and 11.2% for the sqDIVA-T and sqDIVA-S models, respectively. Between the two four-DOF models, sqDIVA-T resulted in a significantly (13.6%, $p=0.00046$) smaller fitting error than sqDIVA-S.

Figure 3.3.C - E show the model fitting results from three representative subjects. These three subjects were chosen to represent the cases in which there were substantial, modest and negligible gains in the fitting quality under the sqDIVA-T models relative to the Baseline model

(indicated by the filled circles in Fig. 3.3.A and B). It can be seen that in the majority of the subjects, the introduction of the timing adjustment into the model led to considerably improved ability of the model to capture the patterns of the F2 compensation profiles. Similar observation can be made for the sqDIVA-S model, despite the fact that the gain of fitting performance was substantially less for the sqDIVA-S (Fig. 3.3.B).

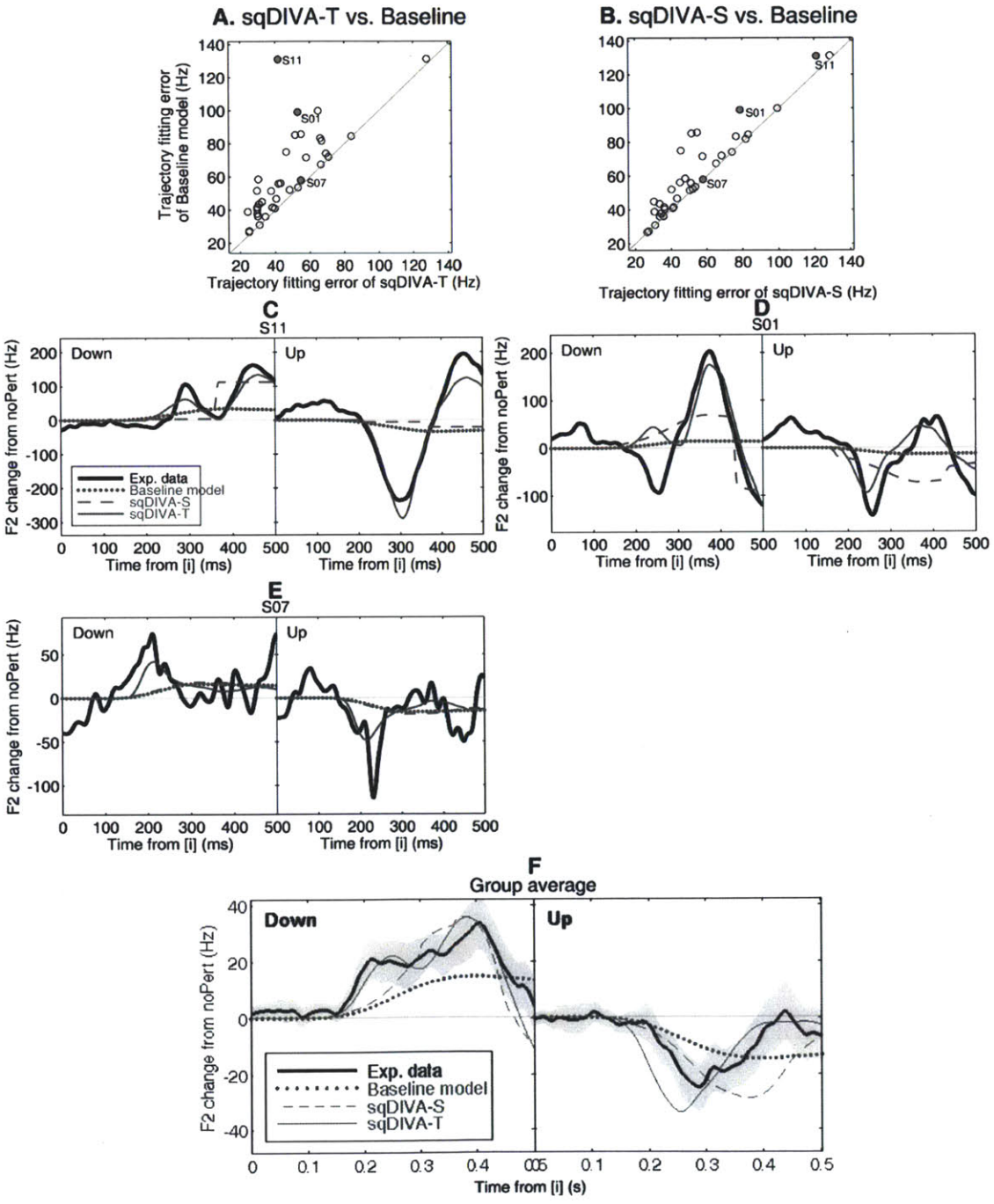


Figure 3.3. (previous page) Performance of the baseline, sqDIVA-S and sqDIVA-T models in fitting the F2 compensation profiles from the Down and Up perturbations. **A:** Comparing the fitting errors of the sqDIVA-T (abscissa) and baseline (ordinate) models. The gray diagonal line indicates equality between the two models. Each circle in the plot corresponds to one individual subject. **B:** Comparing the fitting errors of the sqDIVA-S (abscissa) and baseline (ordinate) models (same format as A). **C – E:** individual-subject fitting results from the three representative subjects. The correspondence of these three chosen subjects with the data points in Panels A and B are indicated by the filled circles in those panels. **F:** The grand average fitting result from the three models compared with the experimental data.

The best-fitting simulated F2 compensation profiles for the individual subjects were averaged to give rise to the average simulated F2 compensation profiles. This was done for all three models. From Fig. 3.3.F, it can be seen that the results from the sqDIVA-T and sqDIVA-S models provided more accurate approximations of the experimental results than that from the Baseline model. Of these two non-Baseline models, sqDIVA-T (thin solid curve) generated a fairly accurate approximation of the group-average compensation profiles from the experiment and its performance was slightly better than the sqDIVA-S model (dashed curve).

Since these gains in fitting the experimental of the sqDIVA-T and sqDIVA-S models may be trivially due to increased DOF relative to the Baseline model, it is necessary to evaluate the performance of these non-Baseline models in ways that are not a part the cost function minimized in the optimization procedure (i.e., the total RMS error of fitting the individual F2 compensation curves). To this end, we chose the errors of the models in fitting the changes in the two major time intervals ([i]-[u]$_1$ and [i]-[j]$_1$) intervals as the measure for this independent evaluation. As can be seen in Fig. 3.4.A, the sqDIVA-T model was significantly more accurate in predicting these time interval changes under the Down/Up perturbations than the Baseline model ($p=0.0014$). On average, the error reduction of the sqDIVA-T model was 21.2% relative to the Baseline model. By contrast, the sqDIVA-S model, which had the same number of DOF as the sqDIVA-S model but didn't incorporate timing adjustments, failed to show any significant gain of performance in fitting the time-interval corrections. In fact, its performance was even slightly worse (1.9%) than that of the Baseline model. The performance of fitting the time-interval

changes was significant better under the sqDIVA-T model than under the sqDIVA-S model (22.8% difference, p=0.001) (see Fig. 3.4.B). The superior performance of the sqDIVA mode in fitting the time-interval changes can also be seen from the average simulated time intervals (Fig. 3.4.C and D). The average changes in the [i]-[u]$_1$ and [i]-[j]$_1$ interval simulated by the sqDIVA-T model approximated the average changes in the experimental data any substantially more accurately than the results from both the sqDIVA-S and Baseline models.

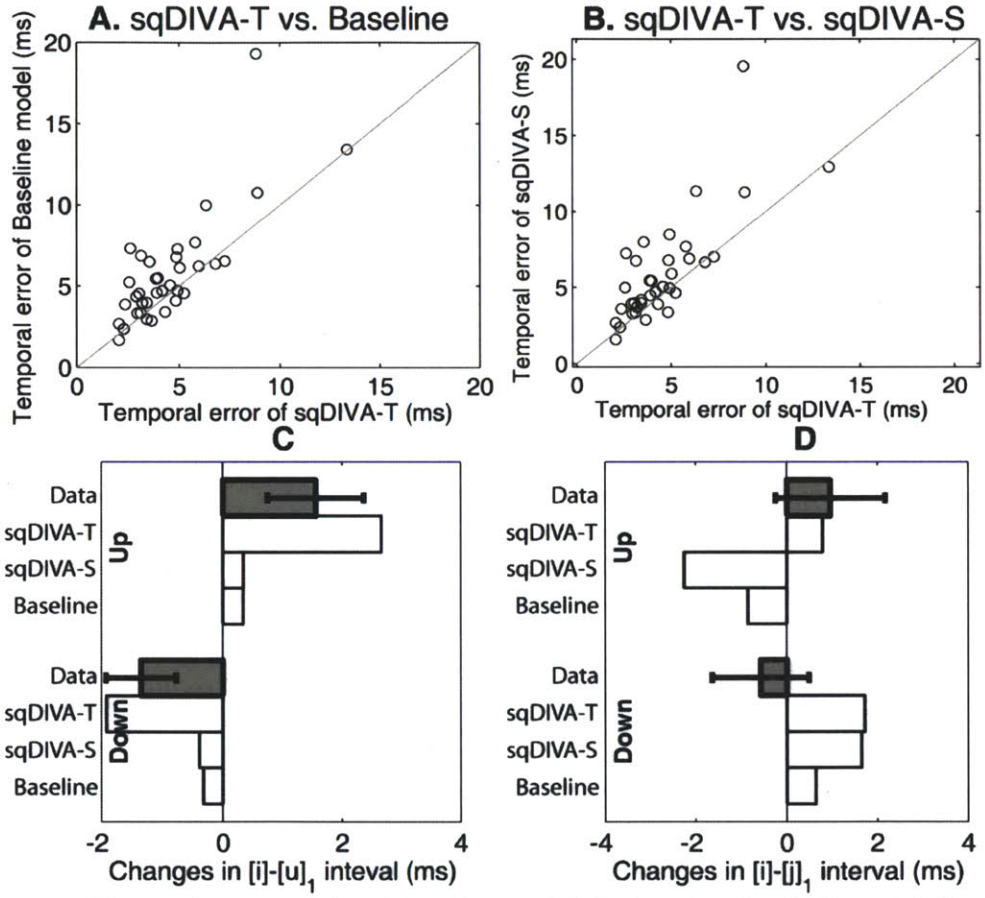


Figure 3.4. The performance of the baseline, sqDIVA-S and sqDIVA-T models in predicting the timing corrections under the Down and Up perturbations. **A:** Comparing the temporal fitting errors of thte sqDIVA-T (abscissa) and baseline (ordinate) models. The gray diagonal line indicates equality between the two models. Each circle corresponds to one individual subject. **B:** Comparing the temporal fitting errors of sqDIVA-T (abscissa) and sqDIVA-S (ordinate). Same format as A. **C:** Performance of the three models in predicting the changes in the [i]-[u]$_1$ interval under the Down (top half) and Up (bottom half) perturbations. The horizontal and vertical error bars show ±1 SEM. **D:** same as C, but for the fitting of the changes in the [i]-[j]$_1$ interval.

From these observations, we can conclude that the accuracy gain for sqDIVA-T relative to the Baseline model was not simply a consequence of the increased number of DOFs. The other four-parameter model, namely sqDIVA-S, which had the same number of DOFs as sqDIVA but possessed a spatial, rather than temporal correction option, was not as good as sqDIVA-S model in capturing the spatial or temporal features of the experimentally observed production compensation.

There are two revelations from these modeling findings. First, the timing corrections observed in the experimental data were not merely byproducts of a control process attends to only spatial parameters of articulation (as embodied by our Baseline model and the current DIVA model). Second, these results provide support for the notion that online AF-based correction of timing exists in the speech motor system during the articulation of multisyllabic, running speech.

3.3.2. Modeling of the Accel and Decel perturbations

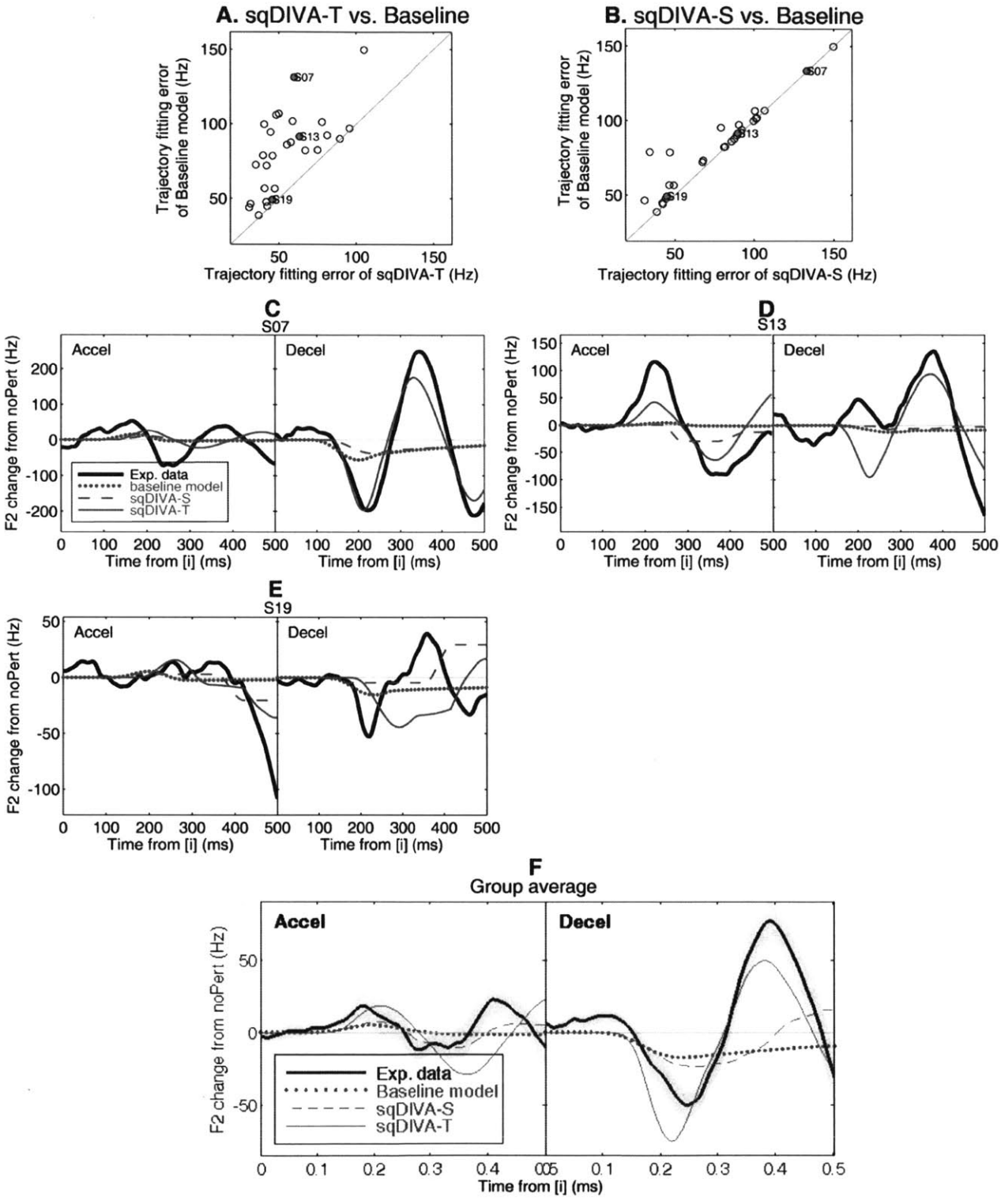
As for the simulation of the Up/Down perturbation, both the sqDIVA-T and sqDIVA-S models showed better fitting of the F2 compensation profiles than the Baseline model. Figure 3.5.A shows that the decrease in the fitting error from the Baseline model to sqDIVA-T was substantial for most of the 28 subjects (Mean: 32.1%, $p < 1\times10^{-6}$, t-test). By contrast, the sqDIVA-S model showed only modest reductions in fitting error relative to the Baseline model (Fig. 3.5.B, mean: 7.2%, $p = 0.0053$, t-test).

Example individual-subject results are shown in Panels C, D and E of Figure 3.5. Panel C shows a subject for which the sqDIVA-T model (the thin solid curve) lead to a large amount of error reduction compared to the Baseline model. In this subject, the sqDIVA-S model produced a fitting result comparable to that of the Baseline model, i.e., much less accurate than the sqDIVA-T model. Panel D shows a case in which the performance gain afforded by the sqDIVA-T model

relative to the Baseline model was intermediate. As in Panel C, the sqDIVA-S model did not lead to any noticeable improvement of fitting either. Panel E shows a subject for which the performances of the three models (sqDIVA-T, sqDIVA-S and Baseline) are similar to each other, i.e., a case in which substantial fitting error reduction was seen under neither sqDIVA-T nor sqDIVA-S.

For each of the three models, the optimally fitted curves from the individual subjects were averaged to generate the simulate group-average compensation curves (Fig. 3.5.F). Both the Baseline and the sqDIVA-S models produced average curves that are very dissimilar to the experimental data in both the shape and the magnitude of the curves. In comparison, the average curve from sqDIVA-T was much more similar to the experimental data, in shape (such as the number and approximate timing of the inflection points) and magnitude. The fitting performance of the sqDIVA-T model was not perfect, as the magnitude of the compensation in its group-average curve fell out of the mean±1 SEM bound (shaded region) at many time points, but is still clearly superior to those of the Baseline and sqDIVA-T models. Therefore it can be seen that sqDIVA-T outperformed sqDIVA-S by a considerable degree, despite the fact that both models have the same number of DOFs, which attests to the superiority of the principle of feedback-based online timing adjustment.

Figure 3.5. (next page) Performance of the sqDIVA-T, sqDIVA-S and Baseline models in fitting the F2 compensation profiles under the Accel and Decel perturbations. **A:** Comparing the fitting errors of the sqDIVA-T (abscissa) and baseline (ordinate) models. The gray diagonal line indicates equality between the two models. Each circle in the plot corresponds to one individual subject. **B:** Comparing the fitting errors of the sqDIVA-S (abscissa) and baseline (ordinate) models (same format as A). **C – E:** individual-subject fitting results from the three representative subjects. The correspondence of these three chosen subjects with the data points in Panels A and B are indicated by the filled circles in those panels. **F:** The grand average fitting result from the three models compared with the experimental data. The experimental data are shown by the thick solid curves (mean) and the shaded regions (±1 SEM).


A. sqDIVA-T vs. Baseline
Trajectory fitting error of Baseline model (Hz)
Trajectory fitting error of sqDIVA-T (Hz)
S07
S13
S19
B. sqDIVA-S vs. Baseline
Trajectory fitting error of Baseline model (Hz)
Trajectory fitting error of sqDIVA-S (Hz)
C
S07
Accel
Decel
F2 change from noPert (Hz)
Exp. data
baseline model
sqDIVA-S
sqDIVA-T
Time from [i] (ms)
D
S13
E
S19
F
Group average
Accel
Decel
Exp. data
Baseline model
sqDIVA-S
sqDIVA-T
Time from [i] (ms)

The performances of the models in predicting the timing adjustment patterns under the Accel and Decel perturbations are shown in Figure 3.6. Panel A compares the timing-adjustment fitting errors of the sqDIVA-T (abscissa) and Baseline (ordinate) models, from which we can see the reduction of fitting errors in vast majority of the 28 subjects. The mean timing prediction error reduction of sqDIVA-T relative to the Baseline model was 28.4%, which was significantly different from zero ($p<0.00005$, paired t-test). The sqDIVA-T not only showed performance better than the Baseline model, but also outperformed the sqDIVA-S model. As Fig. 3.6.B shows, for most of the subjects the timing-adjustment prediction error was smaller under sqDIVA-T than under sqDIVA-S (mean difference: 28.1%, $p<0.00005$, paired t-test).

Panels C and D of Fig. 3.6. focuses on the details of the timing change predictions. As can be seen in these panels, sqDIVA-T is the only one of the three tested models that was capable of accurately predicting the respective shortening and lengthening of the [i]-$[u]_1$ and [i]-$[j]_1$ intervals under the Accel and Decel perturbations. The sqDIVA-T model's predictions of the changes in these two time intervals were not only qualitatively accurate, but all predictions on average were within the bounds of mean±1 SEM, indicating the quantitative accuracy of the model. By contrast, both the Baseline and sqDIVA-T models predicted timing adjustments close to zero and much smaller than the experimentally observed values. These results, together with the results of the Up/Down simulation from the previous section, corroborate our hypothesis that feedback-mediated online timing correction does exist in the speech motor system and it is an important factor in understanding the behavior of this system during the control of multisyllabic articulation.

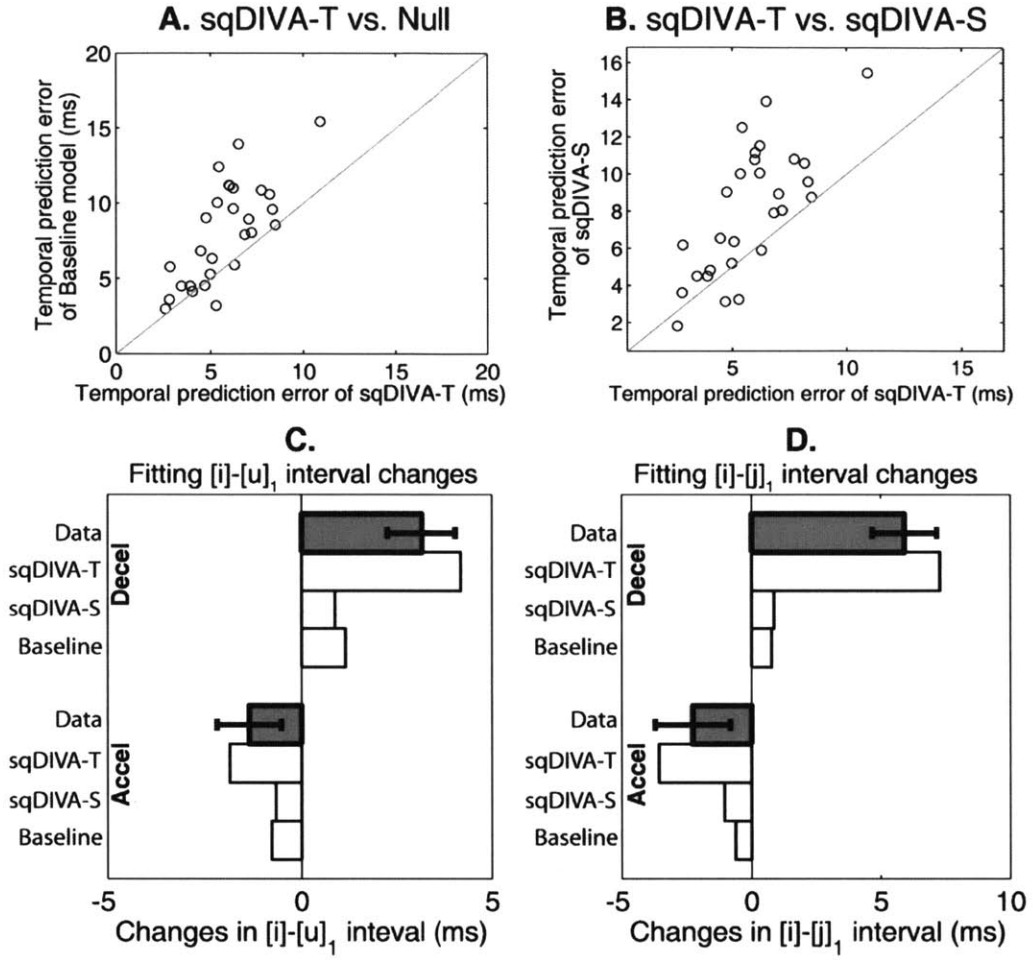


Figure 3.6. Performance of the sqDIVA-T, sqDIVA-S and Baseline models in fitting the timing adjustments under the Accel and Decel perturbations. **A:** Comparing the temporal fitting errors of thte sqDIVA-T (abscissa) and baseline (ordinate) models. The gray diagonal line indicates equality between the two models. Each circle corresponds to one individual subject. **B:** Comparing the temporal fitting errors of sqDIVA-T (abscissa) and sqDIVA-S (ordinate). Same format as A. **C:** Performance of the three models in predicting the changes in the $[i]$-$[u]_1$ interval under the Down (top half) and Up (bottom half) perturbations. The horizontal and vertical error bars show ±1 SEM. **D:** same as C, but for the fitting of the changes in the $[i]$-$[j]_1$ interval.

3.4. Discussion

In this chapter, a computational model of the auditory feedback-based online articulatory control was created and tested as a way to explain the experimental data and to integrate them into a more coherent understanding of the speech motor control process. This computational modeling endeavor was partly motivated by the complexity of the data we were faced with.

Through comparing three variants of the model (baseline, sqDIVA-S and sqDIVA-S), we reached the following conclusions. First, a simple, closed-loop control scheme that doesn't explicitly model the sequencing of multiple articulatory units was inadequate for modeling the spatiotemporal details of the experimentally observed compensatory responses. This observation reinforces the idea that fluent speech production is a sequential process that involves the generation of longer utterances by stringing shorter units of control together. Second, by comparing the sqDIVA-T and sqDIVA-S variants of the model, we established that the data can be more accurately explained by a control process in which the brain constantly uses information from AF to adjust the timing of the articulatory units on a moment-by-moment basis.

Mechanical perturbations to articulation during multisyllabic speech was shown to elicit online compensatory adjustments in both the spatial coordinates and timing of articulatory movements (e.g., Gracco and Abbs 1989; Munhall et al. 1994). It is conceivable that the speech motor system uses somatosensory feedback in conjunction auditory feedback to control the timing of articulatory movements on a moment-by-moment basis. The sqDIVA-T model outlined in this chapter focused on auditory feedback in order to demonstrate the principal of such online timing control; future iterations of the sqDIVA model will incorporate somatosensory feedback as control signals. The feedback-based timing control in sqDIVA-T is somewhat similar to the model of Kalveram (1987), which is based on premised that AF is used by the speech motor system to control syllable timing, in at least stressed syllables. However, sqDIVA-T differs from Kalveram's (1987) model in at least two marjos respects. First, sqDIVA-T is concerned with not only the timing of syllables, but also the detailed spatial trajectories of the articulatory movements during the production of the syllables. Second, sqDIVA-T postulates that AF-based timing control is active during both stressed and unstressed syllables, whereas Kalveram's model

posits that AF-based timing control functions during only stressed ones. As for the latter difference, the experimental evidence in this thesis is not capable of supporting either side, because the AF perturbation was focused on monosyllabic words.

It is noteworthy that the timing change patterns observed by Gracco and Abbs (1989) are qualitatively consistent with the formulations of the sqDIVA-T model. They applied a downward force load to the lower lip during the production of the pseudoword "sappaple" and observed that the onset of the second [p] closure movement and the associated muscle activity occurred earlier under this perturbation. This timing shift can be explained by the principle of the sqDIVA-T model: the perceived overshoot caused by greater-than-unperturbed velocity of the lip lowering during the second syllable [pæ] may have triggered the earlier-than-normal onset of the end of the syllable (marked by the onset of the [p] closure movement) under the downward force perturbation.

In the sqDIVA-T model, two modes, simple and complex, of feedback-based correction coexist. The simple feedback based correction doesn't depend on the sequential nature of the motor task, and is similar to the control process observed through perturbation experiments on monophthongs. By contrast, the complex process is closely related to the sequential nature of the motor program. It modifies the feedforward motor programs, which leads to preemptive minimization of speech motor errors. It is likely that these two control processes are based on distinct neural substrates. Tourville et al. (2008) observed the involvement of bilateral posterior superior temporal cortical areas and right motor and premotor cortices in the simple-type online correction. Based on previous findings, we postulate that the basal ganglia and cerebellum are involved in the complex, timing correction process, as these subcortical structures have been shown to participate in timing control of speech (Wildgruber et al. 2001; Ackermann 2008) and

lesions in these structures lead to loss of normal speech timing in speech disorders such as ataxic dysarthria (Spencer and Slocomb 2007). This hypothesis awaits testing with future fMRI studies.

The "timing score" in the sqDIVA model plays a role similar to that of "gestural activation coordinates" in the Task Dynamic (TD) model (Saltzman, 1989). Both the timing scores of sqDIVA and the activation coordinates of TD are generated largely by hand, i.e., in ways that are consistent with the basic premise of the extrinsic timing theory of Fowler (1980). They are both agnostic about the underlying mechanism that generate the timing patterns in the first place (but see more recent work of the Haskins group to suggest coupled oscillators as a possible bases for some aspects of the activation coordinates, e.g., Saltzman et al. 2006). However, the timing scores also differ from the gestural activation coordinates in two aspects. First, as a strength of the sqDIVA model, the timing scores of sqDIVA interacts with sensory feedback-based control mechanisms, unlike the activation coordinates in TD, which are feedforward processes that do not incorporate mechanisms for online feedback-based adjustments. Second, as a weakness of the sqDIVA model, gestural activation coordinates in TD are multidimensional as they correspond to multiple tract variables; the activation "waves" for different tract variables in TD are allowed to overlap in time, which offers a way of modeling coarticulation phenomena in speech. By contrast, currently the timing scores in sqDIVA are one-dimensional (only for F2) and do not permit the temporal overlapping of gestures as the TD model does. This simplification of the sqDIVA model allows us to focus on the key issue the model aims to address, namely the role of AF in the online control of intergestural timing. Future iterations of the model will increase the dimensionality of the controlled states and that of the timing score. The principles of online feedback control we developed for this one-dimensional sqDIVA will be adapted with relatively minor modifications to accommodate the increased dimensionality.

Apart from the DIVA model, another existing model that explores the interactions between auditory-motor interaction in the speech system is the State Feedback Control (SFC) model (Ventura et al. 2009; Hickok et al. 2011). In the SFC model, AF is used by the speech motor system to update an internal representation of the current state of vocal tract, which is in turn used by an internal forward model to make predictions of the future auditory feedback. The predicted future AF is compared with the auditory target of the sound being produced to update speech motor commands. The sqDIVA model is similar to the SFC model in important ways, including the look-ahead manner in which AF is utilized by the speech motor system. But sqDIVA also differs from SFC in certain respects, including the fact that sqDIVA utilizes a pre-learned (yet online-adjustable) set of motor commands rather than computing those commands on the fly. In addition, the mathematical details of the SFC models have not been established yet, therefore this model has a "boxes-and-arrows" formulation that cannot generate quantitative, testable predictions. In comparison, the sqDIVA model is mathematically explicit and hence is capable of generating concrete and testable predictions.

One important limitation of the current study is the choice of the stimulus utterance, which was atypical to a certain degree because it comprised of only vowels and semivowels. This choice of stimulus utterance was due to a specific methodological consideration: our AF perturbation system was devised to function mainly on vowel-like speech sounds, not on consonants such as stops. Future studies will be needed to confirm the existence of the same spatial and temporal correction processes during the articulation of more generic types of multisyllabic utterances that contain those types of consonants.

The sqDIVA-T model is aimed at illustrating the interaction between AF and motor timing and sequencing in running speech. Admittedly, the sqDIVA-T model is a theoretical framework

that is purposefully kept simple by ignoring many aspects of speech articulation, such as articulator trajectories and articulatory-acoustic relations. Efforts will be made in the future to incorporate the control principles of sqDIVA-T model into the framework of the DIVA model to overcome this limitation, focusing on the sensorimotor process underlying the timing control in multisyllabic articulation, sqDIVA-T model is located on an intermediate level and will be a key piece for a successful integration of GODIVA, a cognitive model with no sensorimotor processes (Bohland et al. 2009), and DIVA, a sensorimotor model that is concerned with single-syllable articulation. Despite being significantly better than the alternative models in predicting the magnitude and timing properties of the compensatory responses, the predictions of sqDIVA-T model still differ from the experimental data by considerable amounts. Clearly there is substantial room for improvement relative to the performance of sqDIVA-T. Some improvements may be realized through modeling details of the articulatory process by using model vocal tracts; additional improvements may be afforded through examining the role of other formants left out in the current simulation, e.g., the first formant trajectory. Despite being imperfect, the sqDIVA-T model is a first attempt to address the sensorimotor processes in fluent running articulation and sets ground for future experimental and theoretical work on this complex albeit important issue.

Chapter 4. Auditory feedback and online feedback control of articulation in stuttering

4.1. Introduction

As we have discussed in the literature review in Section 1.2.2, several theories and models about the etiology of stuttering have focused on the function of internal models and the role of AF in this disorder (Max et al. 2004; Civier et al. 2010; Hickok et al. 2011). These models agree in that AF is used in certain abnormal ways by the speech motor system of a PWS, and these abnormalities in the usage of sensory feedback are related to the breakdown of speech fluency in stuttering. However, these models differ in which part of the speech motor system is hypothesized to be abnormal and how AF is utilized in abnormal ways. Based on the review in Section 1.3.2, we can categorize the theories into two broad categories.

The first category, which we refer to as the *over-reliance theories*, is exemplified by Civier (2010) and Hypothesis 2 of Max et al. (2004). These theories proposed an over-reliance on AF in speech motor control, which is possibly a consequence of a defective feedforward pathway related to white-matter deficits of the brain (c.f., Sommer et al. 2002; Chang et al. 2008; Watkins et al. 2008; Cykowski et al. 2010). The second category, which we refer to as the defective internal model theories, has been proposed in several previous papers. For example, Hypothesis 1 of Max et al. (2004) and the model by Hickok and colleagues (2011) both propose that noisy or unstable internal models (including forward and inverse models that translate parameters between the motor and auditory domains) are the key defects of the speech motor system in PWS. The earlier theory of Neilson and Neilson (1987) is similar to the two above-mentioned models but differs slightly from them in that it only emphasizes defects in the inverse internal model.

Which of the two categories is theory is closer to the reality? An answer to this question can lead to a significant advance in our understanding of the etiology of stuttering. Here I argue that the method of randomized auditory perturbation (see Sect. 1.1.4) can be used as a tool to test

these theories, because the two categories of theories generate predictions that are directly testable within this experimental paradigm.

Randomized perturbations of the first formant frequency (F1) during the production of steady-state English vowel [ɛ] has previously been shown to lead to short-latency, online compensations, which take the form of adjustment of the produced F1 in directions opposite to those of the perturbations. These corrections are small in magnitude but statistically significant (Purcell and Munhall 2006b; Tourville et al. 2008). To my knowledge, no published studies have examined this type of online formant compensation in PWS and how they differ from the compensations made by fluent controls.

The two categories of theories reviewed above generate readily distinguishable predictions about the outcome of such an experiment. First, the over-reliance theory predicts that the compensatory formant adjusts in PWS should be greater in magnitude compared to those of the control subjects. To see why the theory makes such a prediction, we can take a close look at the simulations done by Civier et al. (2010). In order to simulate stuttering-like behavior, Civier et al. (2010) adopted a high feedback weight (α_{FB}) of 0.75 in the DIVA model, considerably greater than the value used in "normal" versions of the model (0.15, Tourville et al. 2008). They showed that this abnormal bias toward feedback control leads to large formant errors in production, which if coupled with a resetting mechanism, may generate defective articulatory movements that are similar to dysfluency events in stutterers. A corollary of this model of stuttering, which doesn't require the additional feature of resetting, is that stutterers should compensate more to unexpected online perturbation of AF than normal speakers do. The reason is as follows: in the DIVA model, a higher feedback weight will lead to a larger compensatory F1 change in response to perturbation of the auditory feedback of F1 (see Equation (1.1)). In order to demonstrate this, we performed computational simulation of F1 perturbation during the production of the steady-state vowel [ɛ] with DIVA with 8 different values of α_{FB}. The magnitude of the perturbation was

fixed at 20% of the original value[23]. As the simulation results in Fig. 4.1. show, there is a monotonic and positive relation between the feedback weight α_{FB} and the magnitude of the compensatory F1 change. Hence if the AF over-reliance model by Civier et al. (2010) is correct, we expect to see greater compensatory formant changes in the PWS group than in the control group. Although the example in Fig. 4.1 uses a static articulatory gesture, the aforementioned positive relationship between the feedback weight and the magnitude of the compensatory adjustments should hold for time-varying articulatory gestures (such as in multisyllabic utterances) as well, because Equation (1.1) is valid regardless of whether the articulatory target is static or time-varying.

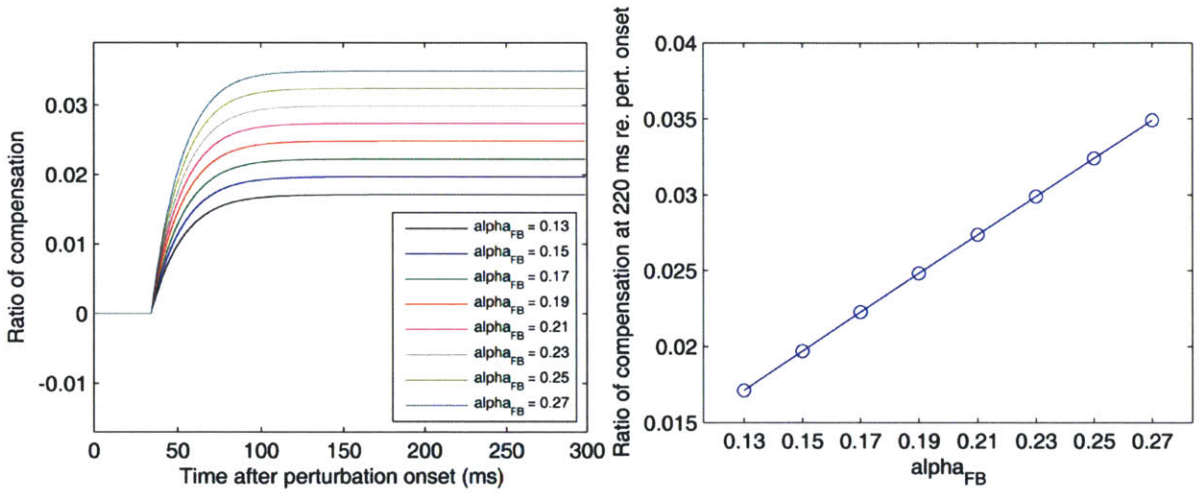


Figure 4.1. Simulation of the relations between the feedback weight α_{FB} of DIVA and the model's compensatory response to perturbation of the AF of F1. Compensatory F1 changes in response to a 20% downward shift of the AF of F are shown1. Left: the F1 trajectories in the model's production under the AF perturbation. Each curves shows the simulation result based on a specific value of the feedback weight (α_{FB}) (see legend). Zero on the time axis corresponds to the onset of the perturbation. As expected, the magnitude of the compensation increases monotonically with increasing feedback weight. Right: the relation between α_{FB} and the ratio of F1 change in the production at 220 ms following the onset of the perturbation.

[23] Further details of the simulation: DIVA version 1.2 was used. The 300-ms-long vowel target had F1, F2, F3 values of 490, 1796 and 2500 Hz, respectively. Before the perturbation test, the vowel target was practiced for 6 times for the learning to stabilize. The perturbation was based on a fixed ratio of 20% downward from the unperturbed F1. Other parameters of the DIVA model used in the simulation: $\varepsilon_{motor} = 0.001$, $\varepsilon_{decay} = 0.95$, WeightFeedBack_sound = 3; WeightFeedback_somatosensory = 0.

Unlike the over-reliance theories, the defective internal model theories do not predict larger compensation magnitudes in PWS, but instead predict smaller and/or more variable compensation magnitudes in PWS than in controls. The rationale for this prediction is as follows. The corrective motor commands are generated by the inverse IMs, which translate auditory errors (i.e., mismatches between AF and auditory targets of the sound being produced) into corrective motor commands that will minimize future errors. According to the theories within this category (c.f. Hypothesis 1 of Max et al. 2004; Hikock et al. 2011), the defective IMs will generate corrective motor commands that are not as effective in correcting the auditory errors as the ones generated by internal models of a normal speech motor system.

To summarize the reasoning, an observation of greater-than-normal formant compensations in response to randomized formant perturbation in PWS would be consistent with the over-reliance theories; but if smaller- and/or more variable-than-normal compensations are found in PWS, the defective IM theories will be supported. Finally, if no differences are found between the two groups in the compensatory responses, neither category of theories will be supported, and alternative theories regarding the nature of stuttering will need to be sought. These three possibilities results are not only mutually exclusive but also exhaustive of all reasonable possibilities. Through this kind of "pitching against each other", this experiment constitutes a "strong interference", which is regarded by many authors (Platt 1964; Ajemian and Hogan 2010) as an effective and informative way for guiding the development and revision of scientific theories.

To comprehensively test this hypothesis, we performed a study which consisted of three experiments, which separately examined three different aspects of online AF-based control in speech articulation. In the first experiment, which we shall refer to as **Experiment A** (Sect. 4.2), we used an AF perturbation technique similar to that of Tourville et al. (2008) to examine the utilization of AF by the speech motor system of a PWS in controlling static articulatory gestures and how that differed from normal. In the second experiment (Sect. 4.3), referred to as **Experiment B**, we performed the spatial time-varying perturbation experiment same as

described in Section 2.1 to study the role of AF in controlling the spatial (amplitude) parameters of connected speech articulation in stuttering. In the third experiment (Sect. 4.4), referred to as **Experiment C**, we used the temporal perturbation technique as described in Sect. 2.2 to characterize the involvement of AF in the control of timing parameters of multisyllabic speech in PWS and how that differed from normal control subjects.

4.2. Experiment I. Auditory feedback-based control of static vowel articulation in stuttering

4.2.1. Methods

4.2.1.1. Subjects

Nineteen PWS (14 male, 5 female, age range: 17.9-47.0, median: 24.8) and 17 fluent (13 male, 4 female, age range: 19.2-42.6, median: 24.4) control subjects (i.e., PFS) participated in this experiment. The age distributions of the two groups were similar ($p>0.7$, two-sample t-test). The PWS were screened by a certified speech-language pathologist, Dr. Deryk Beal and their diagnosis of persistent developmental stuttering was confirmed. The stuttering subjects' self-reported ages of onset of stuttering were all earlier than 8 years (median: 4.5) and thus the vast majority of them were all typical persistent developmental stutterers. The severity of stuttering of the PWS subjects were assessed with Stuttering Severity Instrument version 4 (SSI-4, Riley 2008). The SSI4 scores covered a range from mild to severe (SSI-4 score range: 13 to 43, median: 24). All subjects were native speakers of American English.

This group of 19 PWS showed a mixed history of treatment. Three of them had no prior history of being treated. Of the remaining 16 PWS, four had undergone group or individual treatment programs within one year of the time of the study. Three of the PWS had been trained as speech-language pathologist, but were naïve to the detailed purposes of the study.

The hearing status of all PWS and PFS subjects were screened with monaural pure-tone audiometry. All PFS subjects had pure-tone thresholds less than or equal to 20 dB HL at 0.5, 1, 2 and 4 kHz, i.e., within the normal range in both ears. All PWS, except two, passed the same

hearing screening. Of the two PWS subjects who failed to pass the hearing screening, a female PWS (PWS_F05) failed to meet this criterion in her left ear at two frequencies (thresholds: 35 dB HL at 0.5 kHz and 30 dB HL at 1 kHz); a male PWS (PWS_M05) failed to meet this criterion in the left ear at two frequencies (30 dB HL at 2 kHz and 40 dB HL at 4 kHz) and in the right ear at two frequencies (25 dB HL at 2 kHz and 30 dB HL at 2 kHz). The data from these two subjects with mild hearing loss in one or two ears were not excluded from analysis for the following reason. A psychophysical test was conducted after the online perturbation experiment of each subject to assess the subject's just noticeable difference (JND) of first formant (F1). This JND test was based on the adaptive staircase procedure similar to the one used by Villacorta et al. (2007). Details of this perceptual acuity test can be found in Sect. 4.2.1.5. By using this psychophysical procedure, we found that the F1 JNDs of the two subjects who had mild hearing loss were not worse than the distribution of the JNDs of the PWS and control subjects who had normal hearing. In fact, the F1 JND of the female PWS with mild hearing loss in the left ear (PWS_F05) had one of the smallest F1 JNDs (i.e., highest sensitivity to F1 change) among all the PWS and control subjects. The F1 JND of the male PWS with mild hearing loss (PWS_M05) was close to the median JND of the other subjects (See Fig. 3.2). These observations indicate that despite the fact that the pure-tone thresholds of the two subjects were higher (i.e., worse) than normal, there was no evidence whatsoever the auditory systems of these subjects were incapable of detecting small shifts in the F1 of the vowel [ɛ] as well as normal-hearing subjects were. Therefore we decided not to discard the data from these two subjects from subsequent analysis.

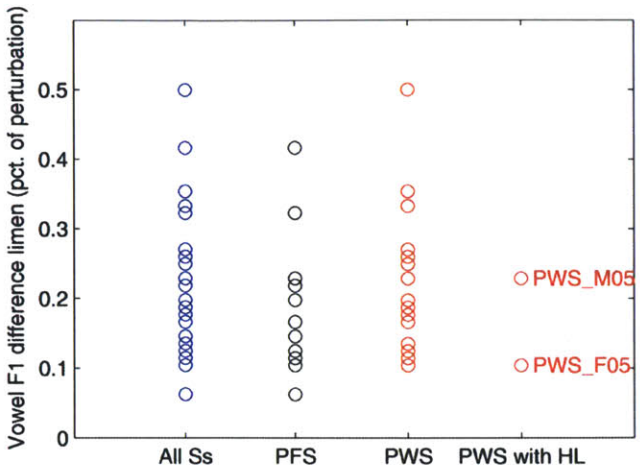


Figure 4.2. Rationale for preserving the data from the two PWS subjects who had mild hearing losses according to our hearing-screening criterion. The blue circles show the distribution of the F1 JNDs (i.e., difference limens) of all normal-hearing subjects, measured with the adaptive staircase procedure. The black and red box-plots show the F1 JND distributions of the normal-hearing PFS and normal-hearing PWS, respectively. The two labeled red circles on the right indicate the F1 JND of the two PWS subjects (PWS_F05 and PWS_M05), which were in the lower range of the JND distribution of normal-hearing subjects, indicating that these two subjects did not have perceptual difficulties in discriminating small differences between the F1s of the vowel [ɛ].

4.2.1.2. Experimental procedure

The auditory perturbation experiment was based on the randomized perturbation paradigm (see Sect. 1.1.4 for definition). Each subject produced the stimulus words "pet" and "head" 80 times each, leading to 160 word-reading trials. These 160 trials were arranged into 20 blocks of eight trials. Each block contained 4 repetitions of each stimulus word, in randomized order. One of the eight trials in a block contained a 20% upward perturbation of F1 (the Up condition); another trial contained a 20% downward perturbation of F1 (the Down condition). The remaining 6 trials contained unperturbed AF (the noPert condition). A constraint was imposed that no two consecutive trials could both contain perturbations. In each block, the two perturbed trials contained different stimulus words. For example, if the Up trial occurred during the word "pet", the Down trial would occur during the word "head". Over the course of the 20 blocks of trials, each word was produced under the Up and Down F1 perturbations for equal number of times, i.e.,

10 times for each perturbation direction. The order of the distribution of the two perturbation directions on the two stimulus words was randomized.

Before the data-gathering part of the experiment (i.e., the 160 trials) began, each subject was trained to produce the words within a normal range of vowel intensity (74 – 84 dB SPL measured by a microphone placed at 10 cm from the subject's mouth) and vowel duration (300 – 500 ms). During the experiment, the onset of offset of the vowel in each trial was identified with an automated procedure based on both short-time root-mean-square (RMS) magnitude of the signal and properties of the spectrogram.

This training was carried out by providing the subject with visual feedback about the success or failure of hitting the aforementioned intensity and duration targets. In the practice phase, the subject is required to repeat a trial until both targets were met simultaneously. In the data-gathering phase of the experiment, the same visual feedback continued to be provided, but a trial was not repeated if the targets were not met. In this way, we ensured that the level and duration of the vowel were consistent across trials, between different perturbation conditions (noPert, Down and Up), and between subjects.

4.2.1.3. Perturbation of formant frequency

The perturbation of the first formant frequency was based on an adapted version of the Audapter software. The formant tracking and shifting algorithms of Audapter has been described in Sect. 2.1.1.3. However, unlike the time-varying perturbation of the formant frequency used in Chapter 2, the perturbation used in this experiment is based on a non-time-varying ratio. Also, unlike the time-varying perturbation experiment, F1, instead of F2, was perturbed in this experiment. The rationale for perturbing F1 is to facilitate comparison with previous studies of online control of formant production, which perturbed F1 of the vowel [ɛ] (Purcell and Munhall 2006b; Tourville et al. 2008).

4.2.1.4. Analysis of the data from the perturbation experiment

The experimenter manually examined all trials from all subjects and identified the ones that contained production errors. These trials were excluded from subsequent data analysis. Trials with production errors amounted to 0.18% of all trials in the PFS group and 0.23% of all trials in the PWS group. Of the trials from the PWS group, 0.066% contained dysfluencies. The repetitive nature of the task led to a low proportion of dysfluent productions by the PWS subjects. Unsurprisingly, none of the trials from the control subjects contained dysfluencies. The algorithm used in automatic detection of the vowel onsets and offsets failed occasionally, resulting in the vowel onset or offset labels being placed too early or too late along the time axis. The experimenter manually corrected the onset and offset of the vowels in these cases of failure of the algorithm. The vowel onset was defined as the onset of the glottal cycles associated with the vowel following the frication noise in the consonant [p] or [h]. The vowel offset was defined as the onset of the F1 decrease leading to the coda consonant [t] or [d]. These definitions of the vowel onset and offset captured parts of the vowel [ɛ] with a relatively constant F1 (see Fig. 3.2 for an example).

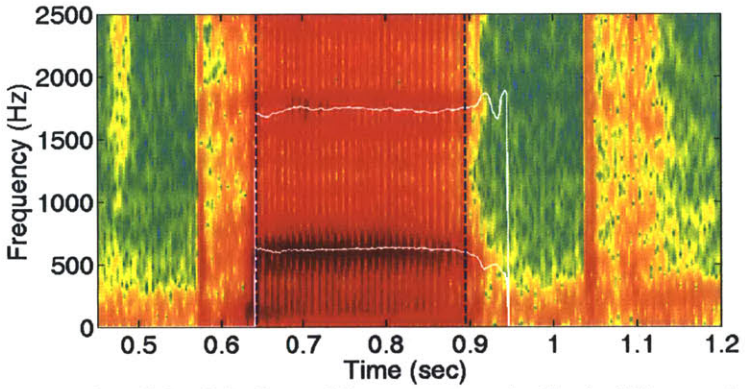


Figure 4.3. An example of the labeling of the onset and offset of the nucleus vowel [ɛ] in the word "pet". The white curves show the F1 and F2 values calculated on the voiced part of the utterance, as detected by RMS thresholding. The two dashed blue lines indicate the onset of offset of the analyzed part of the vowel. See text for details on the definition of the vowel onset and offset. Zero on the time axis corresponds to the beginning of the trial.

Based on the semi-automatically determined vowel onset and offset, an F1 trajectory is extracted from each trial. The F1 trajectory was smoothed with a 28 ms-wide Hamming window. For each subject, the F1 trajectories from the same perturbation status were aligned at the onset

and average along the time axis. The across-trial, within-condition means and SDs were calculated on a frame-by-frame basis. The frame duration was 1.333 ms. This averaging led to three average F1 trajectories from each subject, from the noPert (baseline), Up and Down conditions.

The duration target range for the vowel used in the experiment was [300, 500] ms. The lower bound of this target range intended to ensure that the trials were long enough to engage the online auditory feedback mechanism, which involves a latency around 100 – 200 ms as shown by previous studies (e.g., Burnett et al. 1998, 2002; Larson et al. 2000; Donath et al. 2002; Natke et al. 2003; Chen et al. 2007; Tourville et al. 2008). Here, we analyzed the first 300 ms after the onset of the vowel. Therefore the trials with the vowel parts shorter than 300 ms were discarded, in order to ensure a uniform number of trials from the beginning to the end of the analysis time window. 13.87% of the trials in the PFS group and 11.30% of the trials in the PWS group were excluded from subsequent analysis due to this vowel duration thresholding.

4.2.1.5. Measurement of the auditory acuity to vowel F1 change

Following the afore-described AF perturbation experiment, within the same two-hour experimental session, a psychophysical experiment was conducted to measure the auditory perceptual acuity of the subject to changes in the F1 of the vowel [ɛ]. This perceptual test utilized the adaptive staircase procedure (also known as the adaptive up-down procedure, Levitt 1970).

Our implementation of the adaptive staircase involves a series of two-alternative-forced-choice (2AFC) trials. In each trial, three vowel sounds were played in succession, with the second or the third one different from the first (standard) sound, while the remaining one was identical with the standard[24]. Therefore there were two possible scenarios for each trial: ABA, i.e., second sound different from the standard, and AAB, i.e., the third different from the

[24] The duration of each vowel sound was 300 ms. There was 500 ms gap between two consecutive vowel sounds. Hence the stimulus used in each trial had a total duration of 2400 ms. The F0 of the vowel was equal to the mean F0 of the vowel [ɛ] produced in the unperturbed condition of the AF perturbation experiment. The standard and nonstandard vowels were synthesized with a MATLAB implementation of the Klatt synthesizer (Klatt 1980).

standard. The order of the two scenarios were randomly generated with equal probabilities (0.5). The task of the subject was to judge whether the second or the third sound was different from the standard. The subjects were informed verbally that the purpose of the test was to assess what the smallest differences between two vowel sounds they could detect and therefore they should listen carefully, especially when the difference between the standard and the non-standard was small. They were encouraged to make their best guesses when unsure about the correct choices.

To ensure that the result of the perceptual test is generalizable to the AF perturbation condition, the standard sound (A) was a synthesized steady-state vowel of which the F1 and F2 are equal to the most typical vowel [ɛ] produced by the subject in the unperturbed (noPert) condition in the preceding AF perturbation-production experiment. The most typical trial was determined by plotting the F1 and F2 of the vowels on the 2-dimensional formant space and choosing the one that lay closest to the center of gravity of the data set.

In each run of the adaptive staircase procedure, the B (i.e., non-standard) stimulus had a F1 higher than the A stimulus (standard). The amount of the F1 difference was initially set to the magnitude of the perturbation used in the AF perturbation experiment (20%). A two-down-one-up paradigm was used. If the subject made correct choices in two consecutive trials, the amount of the A-B difference was reduced. Conversely, the A-B difference was increased if a wrong choice was made. Each change in the sign of the increment of the A-B difference constitute a *turn*. The absolute amount of the increment of the A-B difference also changed at each turn. The change amount was initially 25% of the original A-B difference (i.e., 5% of the perturbation magnitude used in the production experiment), and decreased according to a harmonic series of the number of turns ($1/n_{Turns}$). Each staircase was terminated as soon as the sixth turn occurred. The amount of A-B difference at the end of each run was determined as the JND of that staircase. Each subject was administered six staircases, with a 3-4 minute break between the third and fourth. The median of the JNDs form the six staircases was determined as the JND of the subject.

4.2.2. Results

The 20% perturbation to the F1 was imposed on subjects' AF in the perturbation conditions. As the F1 differed between subjects, the perturbation received by different subjects had different absolute magnitudes. To analyze the compensatory response to the perturbation, we calculated the ratio between the online F1 adjustment and the perturbation. By using this ratio-based analysis, we aimed to prevent the differences in the absolute magnitude of the perturbation from becoming a potential confound in the group comparison. However, this approach is based on the assumption that for a given subject, the magnitude of the compensatory F1 is related to the magnitude of the perturbation in a simple, linear fashion. This assumption may not hold, as a previous study reported a nonlinear relation between the magnitude of adaptive formant adjustments and that of AF perturbation (MacDonald et al. 2010)[25]. Therefore it seemed prudent to make sure that the absolute magnitude of the F1 perturbation did not differ significantly between the two groups.

The data showed that this was indeed the case, the average magnitudes of the Down perturbation were 118.86±3.14 Hz and 116.53±3.59 Hz (mean±1 SEM) in the PFS and PWS groups, respectively, which did not differ significantly (t-test: $p > 0.56$). Similarly, the average magnitudes of the Up perturbation were 116.44±4.00 Hz and 113.49±3.45 Hz in the PFS and PWS groups, respectively, the difference between which was not significant either (t-test: $p > 0.50$).

[25] It should be noted MacDonald et al. (2010) examined adaptation to AF perturbation using a sustained perturbation paradigm, whereas the randomized perturbation paradigm was used in the current study (see Sect. 1.1.3). It is unclear to what degree the nonlinear perturbation-compensation relation found by MacDonald and colleagues' finding can be extrapolated to online compensation to randomized perturbations as used in this study. The following analyses were out of precaution.

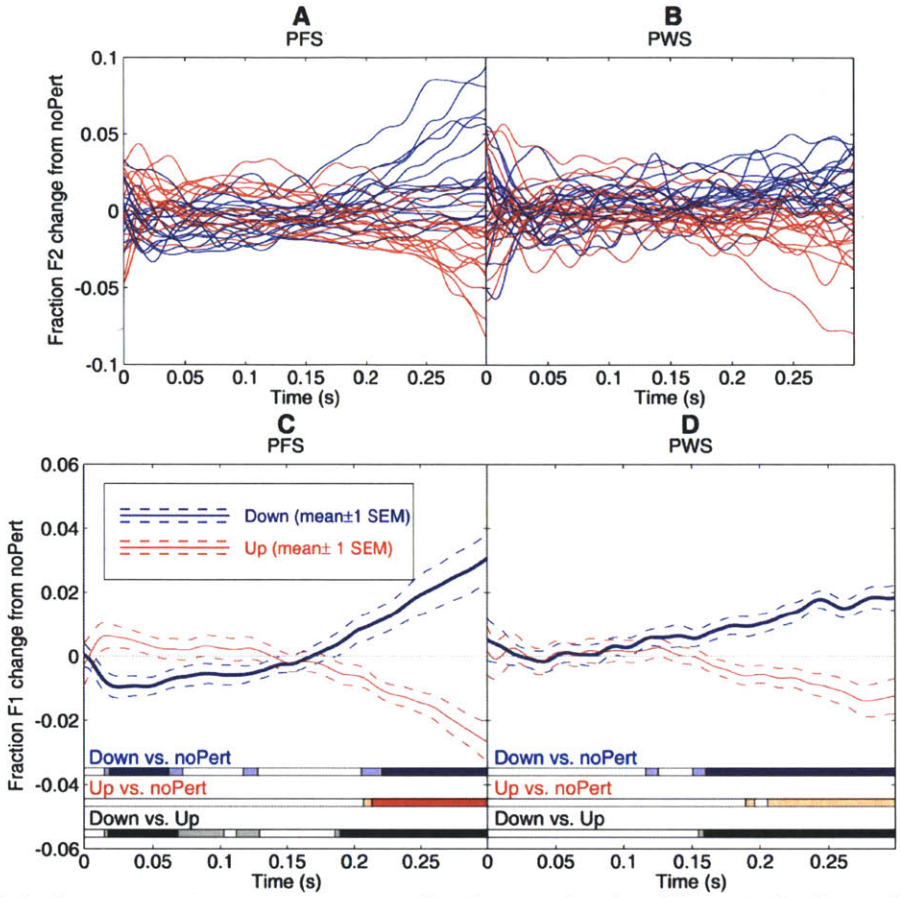


Figure 4.4. Compensatory responses under the randomize F1 perturbations. **A.** Each blue curve shows the average deviations of the F1 trajectories produced under the Down from production under the noPert (baseline) conditions in a PFS subject. Similarly, each red curve shows the response of a PFS subject to the Down perturbation. **B.** The same format as Panel A, but for subjects of the PWS group. **C.** This panel shows summary statistics of the same data as shown in Panel A. Solid curves: average Down-noPert F1 deviation across all 17 PFS subjects; dashed curves: mean±1 SEM. The three horizontal bars on the bottom of this panel indicate significant difference under three comparisons as functions of time. From top to bottom: Down vs. noPert, Up vs. noPert, and Down vs. Up. In each bar, the lighter color (lighter blue, lighter red, or lighter gray) indicate significance at an uncorrected threshold of $p<0.05$. The darker color (e.g., darker blue, darker red, or black) indicate significance at a corrected level of FDR=0.05. **D.** Same format as Panel C, but for the data from the PWS group.

The curves in Panels A and B of Figure 4.4. are the average compensation profiles under the Down (blue) and Up (red) perturbations in the individual subjects. These average compensation profiles were computed by subtracting the average F1 trajectory produced under the noPert condition from that produced under the Down or Up conditions. Panels A and B illustrate the data from the control subjects and PWS groups, respectively. These curves appear rather noisy

and variable across subjects, due to 1) the noise in the individual F1 trajectories and 2) individual variances in the responses made to the AF perturbations.

However, visual inspection of these seemingly noisy curves indicate a common trend in both groups: Starting at approximately 100 - 150 ms following time zero (the onset of the voicing and hence approximately the onset of the AF perturbation) the Down response (blue) curves began to bend upwards, and as consequence, at time 300 ms[26], most of the blue curves in the PFS group were above zero. Conversely, at 300 ms, the majority of the Up response (red) curves had values below zero. These trends can be seen similarly in both the PWS and control groups.

In order to visualize and compare these perturbation-induced changes in the production values of F1 in the PWS and PFS groups, group-average compensation curves were computed by averaging the F1 compensation curves of the individual subjects on a time-frame-by-time-frame basis[27] under each type of perturbations. The group-average from the PFS and PWS subjects are shown in Panels C and D of Fig. 4.4., respectively. The directions and latencies of these curves indicate a compensatory adjustment which commenced at approximately 100-150 ms following vowel onset and became statistically significant at approximately 180 – 220 ms following vowel onset. The magnitude of these adjustments were small, at about 1.5-3% of the baseline (noPert) F1 value, which at 300 ms following vowel onset, approximately accounted for 7.5–15% of the AF perturbation.

To assess the statistical significance of these F1 corrections, we performed statistical analysis of these changes on a frame-by-frame basis. For each subject group, three separate t-test were conducted. For each perturbation condition, the significance of the F1 change from noPert baseline was tested with a one-sample t-test from zero. The frame-by-frame significance results under these tests are shown by the three horizontal bars in the bottom parts of Fig. 4.4.C and D.

---

[26] 300 ms was chosen as the right-side limit of analysis because 1) this was the lower bound of the target vowel duration range used in the experiment and 2) this duration was neither too long ensure the inclusion of the majority of the trials collected during the experiments, nor too short to last beyond the latency of the compensatory responses (~100 – 200 ms).

[27] The length of a time frame used in this analysis was 1.333 ms, due to the same frame duration in the Audapter software (see Sect. 2.1.1.3) for details.

The lighter colors (i.e., lighter blue and lighter red) indicate significance at an uncorrected level ($p<0.05$), whereas the darker colors (blue and red) show significance at a level corrected for multiple comparisons (FDR<0.05, Benjamini and Hochberg 1995). As can be seen from these panels, the PFS group showed significant F1 changes from the noPert baseline with FDR correction under both the Down and Up perturbations. There were also isolated, short periods of uncorrected significant changes early in the trial under the Down perturbation in the control group, although that these early changes were in the directions opposite to the later compensation, i.e., following the direction of the Down perturbation of AF. We will explore the possible cause for this slight early following response in the next section.

The PWS group also showed significant F1 changes from baseline under both the Down and Up perturbations. However, for the Up condition, significant changes were observed only under the uncorrected statistical threshold. As a statistically more sensitive way of confirming the significance of the compensatory responses, matched-sample t-test comparisons were also performed to assess the significance of the difference between the Down and Up conditions. The result of this Down-Up comparison is shown by the bottom-most horizontal bars in Panels C and D, wherein the darker and lighter gray colors indicate significance with and without corrections for multiple comparisons, respectively. As these bars indicate, both groups showed significant differences between the Down and Up condition starting at approximately 160-180 ms following the vowel onset and these differences survived the FDR corrections for multiple comparisons.

These experimental results were largely consistent with the prior data reported by Tourville and colleagues (2008) based on similar speech utterances and similar AF perturbation paradigms. Also, these response patterns appeared to be *qualitatively* similar between the PWS and PFS groups. However, most importantly, the average magnitude of the compensatory responses appears to be smaller in the PWS group than in the control group. Compared to the 3% change at 300 ms after vowel onset in the PFS group, the amount of compensatory change in the PWS group appear to be only approximately 1.5% at the same time point. We will more systematically investigate this between-group difference in compensation magnitude in the following

subsections. But before going into the magnitude analysis, we would like to first briefly explore the cause and implication of the "early following responses" in the PFS group (see Fig. 4.4.C).

Cause of the apparent early following response in the PFS group

We have seen in Fig. 4.4.C that the average trend in the PFS group showed changes in produced F1 in the same direction as the perturbation in early part of the Down- and Up-perturbed trials. What is the cause of this phenomenon? How can this apparent early following response (EFR) in the PWS group have such a short latency? It is prudent to address these questions related to this anomaly before proceeding to further analyses.

To understand the cause of this effect, we need to look at the design of the experiment. Due to the block organization of the experiment, each block contained eight trials (in addition to the filler trial at the end of each block, which we are not concerned with here), six noPert, one Down and one Up trials. The order of these three conditions were randomized, with the constraint that the Down and Up trials must not be consecutive, i.e., they must be separated by at least one noPert trial. Because of this arrangement, within each block, a perturbed trial always occurred after a perturbed trial of the opposite type (if it occurred after any perturbed trials). In other words, it never followed a perturbed trial of the same type in the same block.

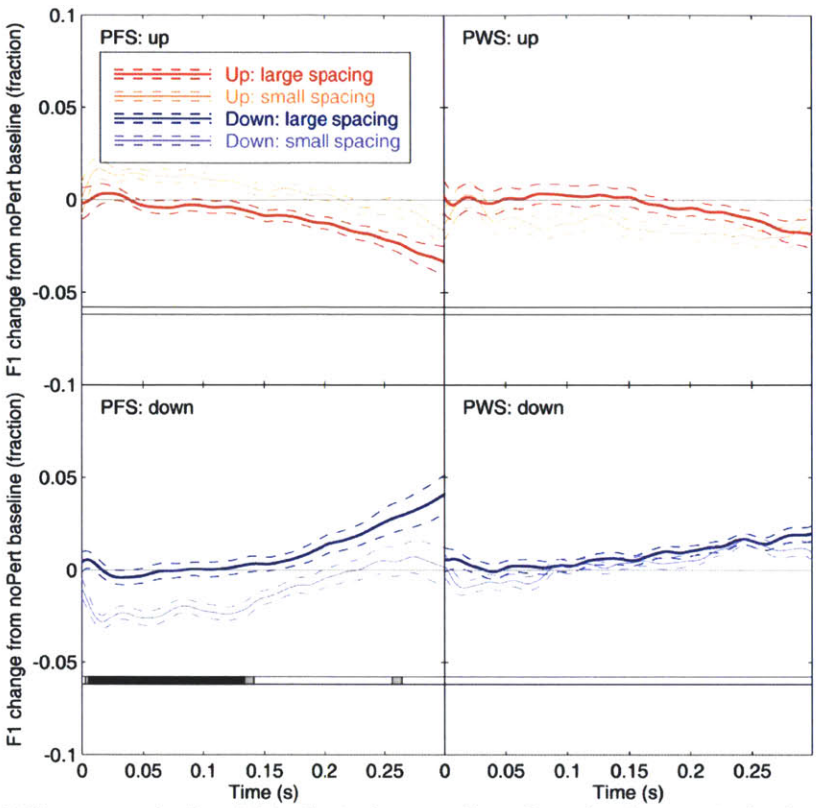


Figure 4.5. Differences in the F1 trajectories produced under the perturbation (Down and Up) conditions with short or long spacing after the preceding perturbation trial. The left and right columns show the data from the PFS and PWS groups, respectively. The top and bottom rows correspond to the Up and Down perturbations, respectively. In each panel, the darker colored (darker red and darker blue) curves are the after F1 changes from the noPert baseline in the perturbed trials separate from the preceding perturbed trial (in the same block) by at least 3 intervening noPert trials or don't following any other perturbed trials in the same block of trials; the lighter colored curves show the average F1 changes from the rest perturbed trials, i.e., those perturbed trials separated from the preceding perturbation trial (in the same bock) by no more than 2 noPert trials. In the PFS group, it is clear that those trials with small spacing showed cross-trial adaptation effects in early part of the vowel. This effect reached statistical significance for the Down condition only. In the PWS group, the same cross-trial adaptation effect could not be observed.

Previous studies have shown that auditory-motor adaptation can occur as following perturbations of AF of F1 (Houde and Jordan 1998; Purcell and Munhall 2006a; Villacorta et al. 2007; MacDonald et al. 2009; Cai et al. 2010). This adaptation effect is different from the online compensation, our main focus in the current study, in that it occurs across trials, rather than within trials. Despite the fact that this experiment is not designed to investigate this cross-trial adaptation, the close juxtaposition of trials in time may lead to some adaptation effects. In fact,

Donath et al. (2002) have reported this type of cross-trial adaptive response in a study based on a randomized F0 perturbation paradigm.

Therefore we hypothesize that if a perturbed trial is sufficiently close to a preceding perturbed trial of the opposite type, the early part of the vowel produced in this trial will show some cross-trial adaptation effect, which may be manifested as a change in the produced F1 in the opposite direction to the AF perturbation in the preceding perturbation trial, or in other words, in the same direction as the AF perturbation in the current trial. It is possible that this cross-trial adaptation effect may be mistaken as the EFR we are investigating.

To test this possibility, we divided the perturbation trials into two categories:

- **Subset A:** the perturbed trials preceded by no other perturbed trials in the same block or separated from the preceding perturbed trial (of the opposite type) by at least 3 noPert trials in the same block, and
- **Subset B**: the perturbed trials that do not fall into subset A, i.e., those separated from the preceding perturbed trial (of the opposite type) by 2 or fewer noPert trials in the same block.

For these two subtypes of perturbed trials, we separately calculated the F1 change from the noPert baseline. The results of this by-subtype calculation are shown in Fig. 4.5. From the left column of this figure, we can see that in the PFS group, there were systematic differences in the F1 trajectories between the type-A and type-B trials. The perturbed trials of type B showed clear cross-trial adaptation effects: The type-B Up trials (the lighter red curve) contained increases in F1 from the noPert baseline in early part of the vowel (<150 ms following vowel onset), such that in the early part of the Up trials, the type-B subset showed higher F1 values than the type-A subset (the darker red curve). A similar cross-trial adaptation effect can be seen under the Down

perturbation in the PFS group (the bottom left panel) and reach statistical significance. In comparison, there was essentially no cross-trial effect of this type in the PWS group[28].

The EFR may become a confounding factor we examine the magnitude of the compensatory response, because it may diminish it by moving in the opposite direction. Therefore in the following analyses, we will adopt a *subset mode* of analysis, in which we will use only the subset A of perturbed trials as defined above. This subset of the trials was less affected by the cross-trial adaptation, as we have shown above. However, because the subset-mode analysis will involve the exclusion of many perturbed trials from the analysis and hence may increase the within- and between-subject variance, we will also preserve the full (non-subset) analysis and show both the non-subset and subset results in the following analyses.

Under the subset mode, 73% of the Down trials and 69% of the Up trials are preserved for the PFS group, and 78% of the Down trials and 73% of the trials for the PWS group[29]. In Figure 4.6., we re-plot the F1 compensation curves as in Fig. 4.4.C and D under the subset-mode analysis. Compared to the non-subset mode, the subset mode diminishes the apparent early following response seen in the PFS before. This indicates that the "early following response" phenomenon was not truly a following response, but a reflection of the unintended cross-trial adaptation effect, which appears to be weaker in the PWS group than in the PFS group.

[28] The results shown here give hint as to weaker auditory-motor adaptation in PWS than in normally fluent speakers, which is an interesting direction of future investigation. This observation is consistent with the slower-than-normal adaptation to sustained AF in adults who stutter compared to controls in the unpublished data by Ludo Max and colleagues (Ludo Max, personal communication). But here we will not dwell on this issue related to auditory-moor adaption further because it is not the main focus of the current study.

[29] These numbers are random variables that cannot be predicted beforehand. This is because the order of the noPert, Down and Up trials are determined randomly in every experiment.

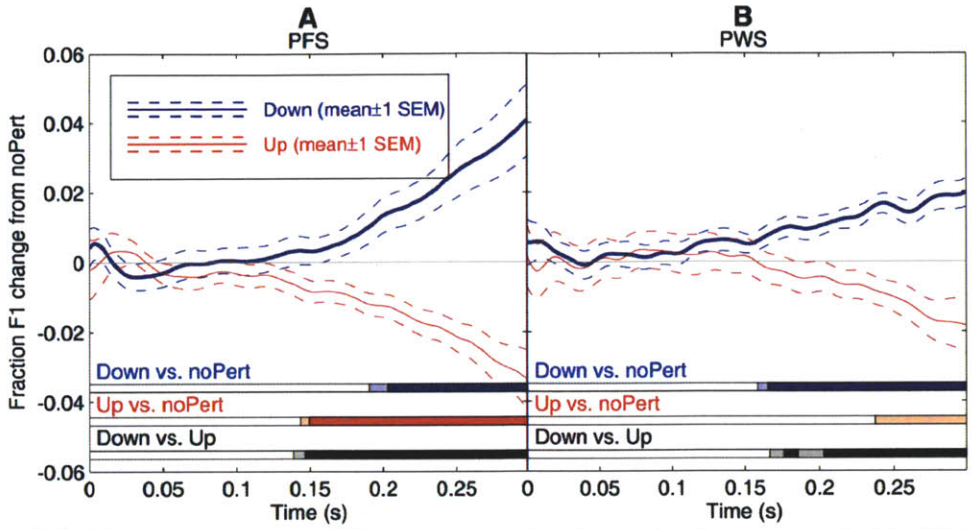


Figure 4.6. The compensatory F1 changes under the subset-mode analysis. The format of these plots are the same as Panels C and D of 4.4. However, they include data from only the perturbed trials of subtype A, i.e., the perturbed trials separated from the preceding perturbed trial in the same block by at least three noPert trials or not preceded by any perturbed trials in the same block. Notice that under this subset mode of analysis, the apparent early following response seen in Fig. 4.4.C is no longer evident. This indicates that the apparent early following response seen before was attributable to a cross-trial adaptation effect (see text for details).

4.2.2.1. Magnitudes of compensatory responses

To summarize the compensatory responses in a succinct way, we computed the *composite response curves* in the two subject groups. In each group, the composite response curve was calculated by subtracting the Up-response curve from the Down-response curve frame by frame. As Fig. 4.7. illustrates, the composite curves of the two groups are similar in shape in that they both show a relatively flat and close-to-zero portion in the first 100-150 ms, followed by a substantial upward bending. Under the subset mode, the curves from the two groups are largely overlapping before 200 ms following perturbation onset. However, the two groups appear to show divergent magnitudes in later parts. Under both the subset and non-subset modes, the two composite response curves of the PWS group showed significantly smaller magnitudes by approximately 48% than that from the PFS group, which indicates weaker compensatory responses to the AF perturbation in stutterers than in nonstutterers.

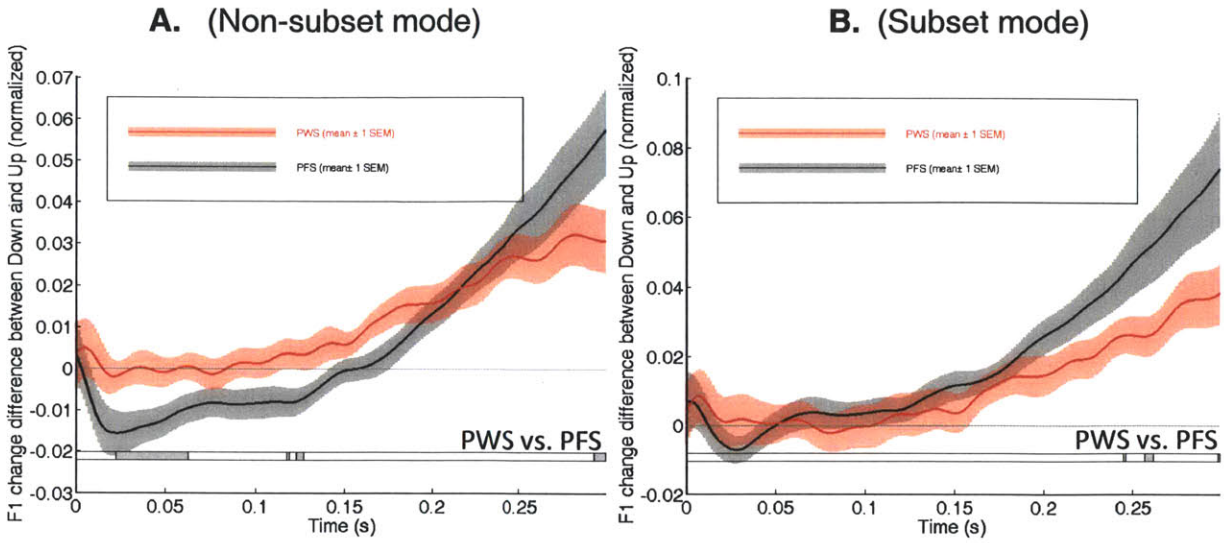


Figure 4.7. Average composite response curves from the PWS (red) and PFS (black) groups. The composite compensation curves were computed by subtracting the F1 change profile under the Up perturbation from the F1 change profile under the Down perturbation. The horizontal bar on the bottom indicates significance of the difference between the two groups on a time-frame-by-time-frame basis (Wilcoxon rank-sum test, two-tailed). Gray: significance on the uncorrected level of $p<0.05$. No significant difference at the corrected level of FDR=0.05 was found at any point from 0 to 300 ms following vowel onset. Panels **A** and **B** illustrate the results under the non-subset and subset modes, respectively.

The between-group comparisons in Fig. 4.7. did not reach significance at a level corrected for multiple comparisons (notice the lack of black color in the horizontal bars in the bottom parts of Fig. 4.7.A and B). To bypass this problem of multiple comparisons, we can choose a single measure is used to quantify the magnitude of the subject's compensatory response. This numerical measure we chose was the value of a subject's composite response curve at 300 ms after vowel onset. Figure 4.8.A and B illustrates the distributions of this response magnitude under the non-subset and subset modes, respectively. Under both types of analysis, the group mean response magnitude based on this measure was smaller in the PWS group than in the control group ($p<0.05$, Wilcoxon rank-sum test, two-tailed).

The results of F-tests indicated there was no significant difference in the between-subject variance of the magnitude of the compensatory response.

To address the question of whether there was any significant correlation between the response magnitude of a PWS and their stuttering severity, we computed linear (Pearson

production moment) and non-parametric Spearman's correlations between the composite response magnitude at 300 ms and their SSI-4 scores. As Fig. 4.8.C and D illustrate, there was no evidence for such a correlation, no matter whether a linear correlation ($p>0.65$) or the non-parametric Spearman's correlation ($p>0.7$) was used. This held true for both the non-subset and subset modes of analysis.

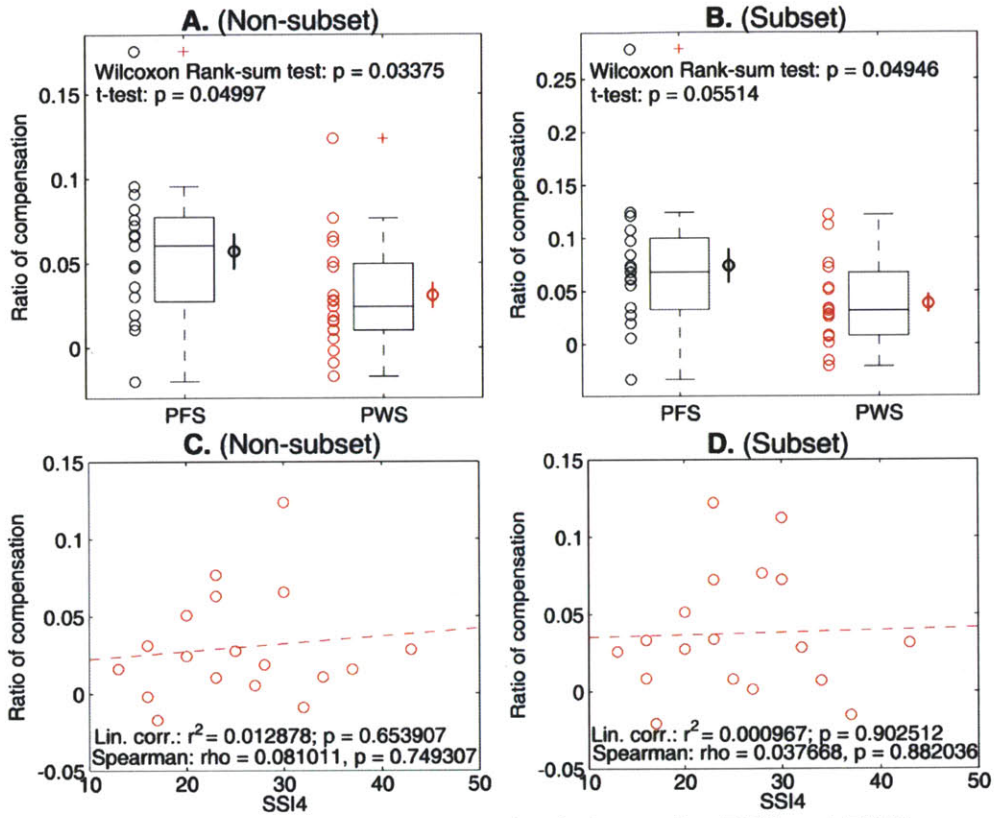


Figure 4.8. Comparison of online compensation between the PFS and PWS groups. **A** and **B:** The left and right columns show the measures of online compensation in the PFS and PWS groups, respectively. Each circle corresponds to one subject. A box-plots is computed based on the data in each subject group, respectively. Panels A and B show the results from the non-subset and subset modes, respectively. **C** and **D**: No significant correlation between the response magnitude and the severity rating score (SSI4). Panels C and D illustrate results from the non-subset and subset modes, respectively.

4.2.2.2. Response latency

It is not appropriate to use the group-average composite response curves as shown in Fig. 4.7. to compute the latency of the compensatory responses, due to the following reasons. First, different subjects may have different response latencies, so the group-average curves, which

smoothes the responses in time, may not be reflective of the latency of any individual subjects. Second, if the group-average responses are used to compute the response latencies, only one measure can be extracted from each group, which will make it impossible to perform statistical comparison between the groups. Therefore, the latency of the response ought to be computed on a subject-by-subject basis.

We used the following custom algorithm to compute the response latencies in single subjects. Suppose the average response curve of the subject under Down and Up perturbation are $r_{Down}(t)$ and $r_{Up}(t)$, respectively, and the across-trial standard deviation of the curve is $\Delta_{Down}(t)$ and $\Delta_{Up}(t)$, respectively. Notice that all the above four quantities are functions of time. We defined the Cohen's d curve $c(t)$, also a function of time, as,

$$c(t) = \frac{r_{Down}(t) - r_{Up}(t)}{\sqrt{\left[(n_{Down}-1)\cdot\Delta_{Down}(t)^2 + (n_{Up}-1)\cdot\Delta_{Up}(t)^2\right] / (n_{Down}+n_{Up})}}$$

$c(t)$ quantifies the effect size of the divergence between the response curves of the Down and Up conditions. Its denominator consists of weighted standard deviations (SDs), rather weighted standard errors. The SDs were was used because it has the merit of being relatively insensitive to sample sizes (i.e., not affected by the number of Down and Up trials). The rationale is that the response latency of a subject should reflect the properties of the subject's speech motor system rather than (partially) reflecting the number of trials in the experiment, which is a completely experimenter-determined extrinsic factor unrelated to the subject's speech motor system. This is reason why we adopted the standard-deviation form in the definition of $c(t)$.

It should be clear that it is meaningful to compute a response latency for a subject only if the subject compensated for the AF perturbation. In other words, if a subject showed no compensatory response to the perturbation, the latency of this subject cannot be defined in a meaningful way. As such, a heuristic criterion for determining if a subject compensated for the perturbation was required. We used the following criterion: for a subject, if $c(300ms) > 0.3$, then

this subject is categorized as compensating[30] (Fig. 4.9.A), otherwise the subject was categorized as non-compensating and his or her response latency was undefined.

For a subject who compensated for the perturbation, we *could* have used the following simple algorithm for determining the response latency: the time at which the Cohen's d curve $c(t)$ crosses the threshold of 0.3 and stays above the threshold until 300 ms could be defined as the latency of the response. However, if defined this way, the compensation latency would be sensitive to the magnitude of response. For example, everything else being equal, a shorter response latency would have been calculated from a subject who showed a greater compensatory F1 changes under the Down and Up perturbations than from a subject who showed a smaller compensatory F1 change, even if the compensatory response in these two subjects started at exactly the same time following the onset of perturbation. Also, this definition would have rendered the latencies sensitive to the variance (noisiness) of the tracked F1s. Since $c(t)$ is inversely related to the SD of the F1, everything else being equal, a shorter response latency would have been calculated from a subject who had a smaller between-trial variance of F1 than from a subject who had a greater variance of F1, even if they actually had equal response magnitudes and onset timing. As such, we should avoid using an absolute-threshold-based approach of computing the latency.

We adopted the following curve-fitting approach for calculating the response latency. The $c(t)$ curve of each subject was fitted with a two-segment line spline $C(t)$ defined by the following equation:

$$C(t) = \begin{cases} x_0 & ,t < L \\ x_0 + b(t-L) & ,t \geq L \end{cases},$$

[30] This threshold value of 0.3 was chosen because conventionally, 0.3 was used as the boundary between "small" effects and "medium" ones in the interpretation of Cohen's d. This selection also took into account practical concerns: smaller values will not be capable of discerning random variation ("noise") from true compensation; and larger (i.e., stricter) ones leave too many subjects without computed latencies. There was some arbitrariness involved in the selection of this threshold. But similar conclusions regarding the latency differences between the PWS and PFS groups were reached if slightly different threshold values were used.

This spline function is comprised of a flat line segment before the latency and a line segment after the latency that has a to-be-fitted slope. It has three free parameters: 1) x_0, the level before the latency, 2) L, the latency (i.e., timing of the turning point), 3) b, the slope of the line segment after the latency. This function was fitted to the $c(t)$ curve of the subject under the least-square-error (LSE) criterion, by which we could determine the value of the latency L.

Under the non-subset mode of analysis, the distribution of the latencies from the 13 compensating subjects out of the 17 subjects in the PFS group is shown by the black bar in Fig. 4.9.B; similarly, the mean latency of the 10 compensating subjects out of the 19 subjects in the PWS group is summarized by the red bar in Fig. 4.9.B. The distributions of the latencies of the two groups were largely overlapping and the difference between the means did not reach statistical significance ($p>0.7$, t-test). Under the subset mode of analysis, slightly more subjects met the criterion for compensation (15 of 17 PFS and 13 of the 19 PWS), but the conclusion regarding the response latencies remained the same: statistically, there was no evidence for a difference in the response latencies of the subjects in the PWS and control groups ($p>0.17$, Fig. 4.9.C).

In the preceding section, we observed a significant difference in the magnitude of the response in the PWS group. One potential concern in interpreting of finding is as follows. It is possible that if one group of subjects have systematically longer latencies of the compensatory response than the other group, then when comparing the response magnitude at a given time following perturbation onset, one would see smaller response magnitude in the former group even if there is no actual between-group difference in the slope of the increase of the response with time (i.e., the magnitude of the response). The lack of evidence for a between group difference in latency alleviates this potential concern.

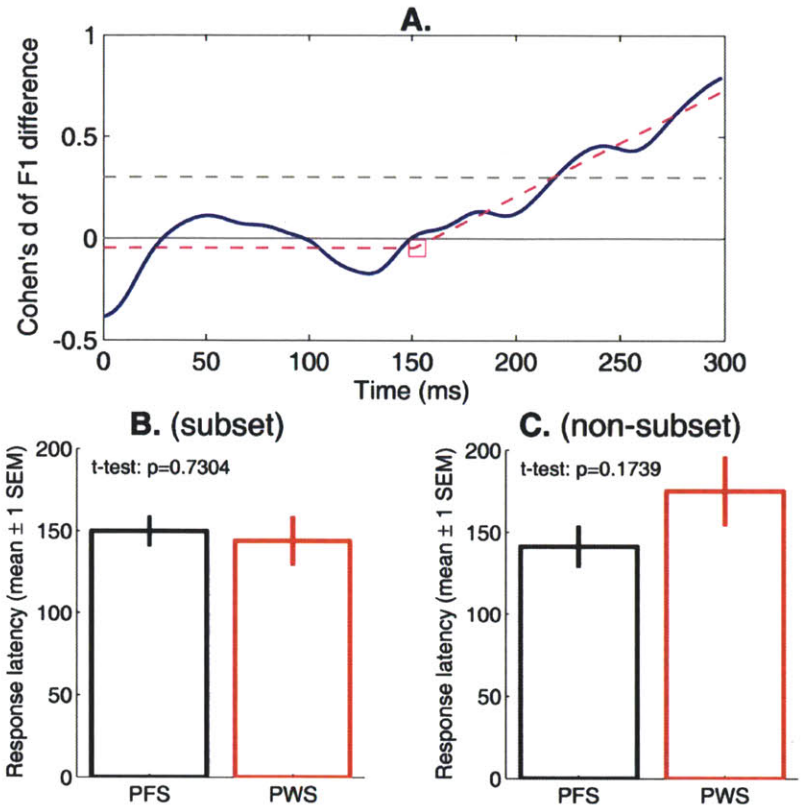


Figure 4.9. Latency of the compensatory responses. **A.** An example showing the way in which the response latency was computed (see text for details). The solid blue curve shows the Cohen's d profile c(t). The horizontal dashed gray line shows the threshold Cohen's d value (0.3). The dashed magenta line shows the two-segment spline fitted to the Cohen's d curve. The magenta square indicates the calculated latency, i.e., the break point of the two spline segments. **B** and **C.** The black and red bars show the response latencies calculated in PFS and PWS groups, respectively. Panels B and C show the results from the non-subset mode and the subset mode, respectively. No significant between-group difference in response latency was found under either mode of analysis.

4.2.2.3. Perceptual acuity to changes in vowel formant in stutterers and nonstutterers

The weaker-than-normal compensation to unanticipated perturbation of the AF of F1 in the PWS group raised the question about possible causes of this under-compensation. A possible cause is that the auditory system of a PWS is less capable of detecting subtle changes in the acoustic properties of speech sounds compared to that of a normally fluent speaker. Several previous studies provided some evidence for abnormal central auditory functions in PWS (e.g., Hall and Jerger 1978; Salmelin et al. 1998; see also review in pp. 184 - 188 of Bloodstein and Ratner 2008). Alternatively, this under-compensation may be attributable to processes other than

auditory processing, e.g., the neural computation (through internal modeling) of proper counteracting motor commands or the implementation of these corrective measures. Without further empirical information, it will be difficult to distinguish between these possibilities.

In the current study, the just noticeable differences (JNDs) of F1 of the subjects were measures as a part of the experiment with an adaptive staircase method (See Sect. 4.2.1.5 for methodological details). In Figure 4.10.A, the JNDs of the two groups are summarized and compared. Along the ordinate, a smaller JND corresponds to a better acuity (higher sensitivity) of the auditory system to changes in F1 of the vowel [ɛ]. The JNDs in Figure 4.10.A are plotted as fractions of the magnitude of the shift (20%) in the perturbation experiment. A JND of 1.0 indicates a magnitude equal to the perturbation used in the AF perturbation experiment. As this figure shows, the JNDs of the PWS were slightly higher than those of the controls (PWS: 0.2135±0.0248; PFS: 0.1667±0.0232, mean±1 SE) but there was no statistically significant difference between the PWS and PFS. In addition, there is no evidence for a negative or positive significant correlation between the JND and the magnitude of the online F1 compensation (Fig. 4.10.B). The results shown in Fig. 4.10.B are obtained under the subset-mode. The results from the non-subset mode are not shown but they are qualitatively consistent with the non-subset-mode results. As such, the findings from the perceptual acuity test do not support the possibility that poorer perceptual ability led to the smaller-than-normal auditory-motor compensation. Therefore the cause of the under-compensation needs to be sought in other parts of the speech motor system, e.g., the internal transformation between the auditory and motor domains. We will further discuss the cause and the interpretation of the auditory-motor under-compensation in Sections 4.4.1. and 4.4.2.

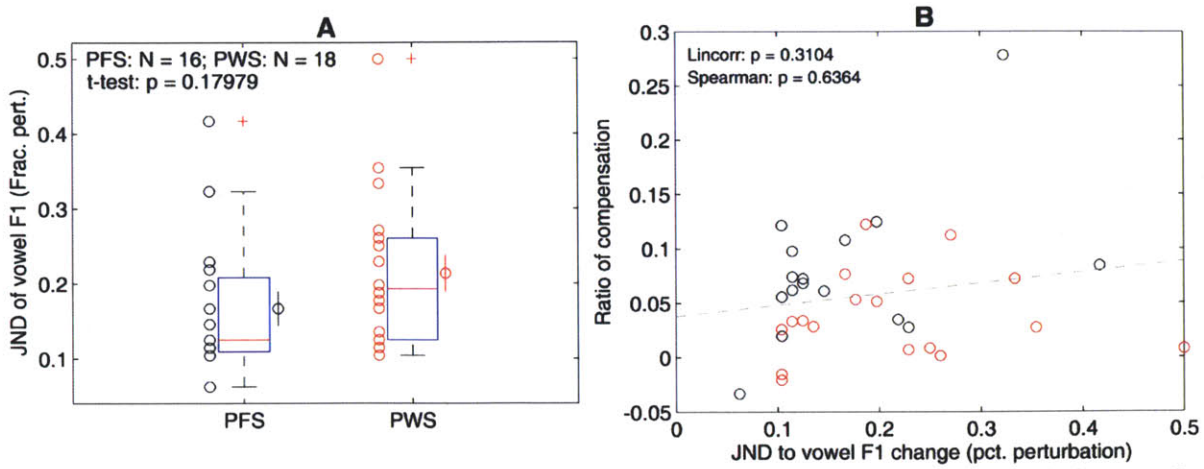


Figure 4.10. Auditory acuity to changes in F1 of the vowel [ɛ] and its relation to the magnitude of the compensation to perturbation in the PWS and PFS groups. A. Comparison of the vowel F1 JNDs between the two groups. A fraction of perturbation of 1 corresponds to the same magnitude of perturbation as used in the production experiment. B. Correlation between the JNDs and the ratio of compensation to the AF perturbation. Compensation magnitudes from the subset-mode analysis is used in this plot.

4.3. Experiment II. The roles of auditory feedback in the online control of time-varying articulation in stutterers and non-stutterers.

In the previous section, we conducted an experiment for comparing the AF-based online control of quasi-static articulatory gestures in PWS and normal speakers. We discovered that although the PWS made online compensatory responses qualitatively similar to normal responses, the magnitude of their compensation were significantly and considerably (48%) smaller than the compensation observed in the normal controls. However, the speaking task used in this static-vowel experiment was unnatural: it involved a holding of a static vowel gesture for an extended period of time (300 – 500 ms). Although this experiment was informative as to the internal auditory-motor transformation performed by the brain in computing simple static articulatory positions needed to achieve a certain acoustic result, the simplicity of the task failed to capture an essential feature of speech production tasks encountered in everyday situations, namely the rapid transitions between sequentially ordered articulatory gestures with appropriate timing patterns. During such sequential, multisyllabic articulation, the brain not only needs to perform the simple static auditory-motor transformation as mentioned above, but is faced with more

complex computational tasks of generating and updating motor commands for transitions between consecutive sounds or syllables, in addition to the adjustment of the timing of success sounds or syllables based on the current state of the vocal apparatus. These online motor command generation and updating activities can utilize AF information, as we have shown in the previous chapter.

Do PWS have deficits in such online time-varying (dynamic) control tasks, in addition to the deficits described above in feedback-related static gestural control? Based on what's already known about stuttering, the answer to this question is likely to be in the affirmative. In PWS, the production of isolated single words can often be quite fluent, while the frequency of dysfluency events increases with increasing complexity of the utterance (Brown 1938; Soderberg 1966; Silverman and Williams 1967). The trial-to-trial variability of articulation (inversely related to the stability of the speech motor execution) is positively related to the length and complexity of the phrase (Kleinow and Smith 2000). Many authors have pointed out that the core deficits in stuttering are centered around the dynamic aspects of speech production (Kent 1984; Ludlow and Loucks 2003). Based on the above experimental results and theoretical considerations, it is important to explore the auditory-motor interaction in dynamic aspects of speech articulation. To our knowledge, this is a question which hasn't been studied to date (but see the somatosensory perturbation study by Caruso et al. 1987 and the preliminary results by Bauer et al. 1997). The experimental methods we developed in Chapter 2 is a suitable tool for addressing this unanswered question.

We conducted the same time-varying AF perturbation experiment on a group of PWS. Their compensatory responses to both the spatial (Down/Up) and temporal (Accel/Decel) perturbations were compared to those of the control subjects. Through this comparative analysis, we discovered that PWS show deficits in the utilization of AF for short-latency online control of the spatial and temporal parameters of articulation in multisyllabic, connected speech.

4.3.1. Methods

Two experiments, called Experiments A and B, which separately examined the online AF-based control of the spatial and temporal parameters of multisyllabic articulation, were conducted on a group of PWS. Two different groups of PFS controls were used in Experiments A and B.

The 18 PWS who participated in Experiments A and B were a subset of the 19 PWS who took part in the static AF perturbation experiment (Experiment I, Sect. 4.2). Useful data could not be obtained from one of the PWS, a 20-year old female, due to the poor formant tracking performance on her voice[31].

Each of the 18 PWS participated in both the spatial and temporal perturbation experiments, in randomized order. Experiments on 8 of the 18 PWS were run under under the S-T (spatial then temporal) order, and the remaining 10 PWS were run under the T-S (temporal then spatial) order. Each PWS subject had a one-to-one matched PFS subject. As described before (Sect. 4.2.1.1), the matching criteria included age (difference < 1 year), gender, and level of education. For each pair of PWS and matched PFS control, the orders of the spatial and temporal experiments were identical. However, in addition to these one-to-one matched PFS, the control groups in Experimens A and B included additional subjects, who were not matched to any of the PWS in a one-to-one fashion.

For the spatial perturbation experiment (A), 36 normally fluent subjects were used as controls. This is the same set of subjects for the spatial part of the study on normal subjects, previously described before in Sect. 2.1.1.1. The ages of these PWS subjects (range: 17.9±47.0, median: 24.9) were not significantly different from that of the PFS group (range: 19.2-42.6, median: 23.8; $p > 0.4$, Wilcoxon rank-sum test). The distributions of the genders did not differ significantly among the PWS and PFS either (PWS: 15M4F; PFS: 30M6F; $p>0.92$, χ^2 test).

[31] The algorithms used for AF perturbation during the multisyllabic utterances in Experiments A and B were more complex than those used for static AF perturbation during the monophthong[ɛ] in Experiment I. A consequence of this higher complexity was a higher sensitivity of the algorithm to noises and irregularities in the formant trajectories, which often resulted from voices that have amodal waveforms, breathiness or other abnormal qualities.

As for the temporal perturbation experiment (B), 28 PFS were used as controls. These were the same subjects as described in Sect. 2.2. The age distributions of the PWS and PFS groups were similar (PWS: range: 17.9–47.0; median: 24.9; PFS: range: 19.2–47.1; median: 24.7) and did not differ significantly ($p>0.75$, Wilcoxon rank-sum test). The gender compositions of the two groups were close (PWS: 15M4F; PFS: 24M4F) and no significant difference was found between the groups ($p> 0.8$, χ^2 test).

Data analysis

To avoid any methodological confounds, analysis of the data from the PWS subjects were based on exactly the same methods as for the PFS subjects, which have been described previously in Section 2.1.1.4.

One analysis-related issue that is unique to PWS is the dysfluencies in the productions. If too many trials contain dysfluencies in the production, the greater-than-normal ratio of discarded trials would cause a concern. However, due to the relative simplicity of the elicitation utterance ("I owe you a yo-yo") and the repetitiveness of the task, which elicited the adaptation effect in stuttering (e.g., Max and Baldwin 2010), the percentages of the trials containing dysfluencies and/or speech errors were quite low in the data from the PWS subjects. In the spatial perturbation experiment, the percentages of trials containing dysfluencies and/or speech errors were 0.21% and 0.26% in the PFS and PWS groups, respectively, which were not significant different between groups ($p>0.1$, Wilcoxon rank-sum test). Similarly, in the temporal perturbation experiment, the percentages were 0.49% and 0.46% in the PFS and PWS groups, respectively, the difference of which was not significant, either ($p>0.6$, Wilcoxon rank-sum test).

4.3.2. Results

4.3.2.1. Experiment A: Responses to spatial perturbation by the people who stutter

On average, the PWS subjects were able to produce the stimulus utterance "I owe you a yo-yo" within the target ranges for duration and speed in 91.3%±11.7% (mean ± 1 SD) of the trials, which did not differ significantly from the same success rate in the PFS group (91.7%±10.7%, $p > 0.85$, t-test). Before analyzing and comparing the compensatory responses of the two groups, it was necessary to ascertain that the magnitude of the F2 perturbations in the AF was similar between the two group, because a difference in the size of the perturbation may lead to differences in the magnitudes of the compensatory responses, even if there are no real differences in the compensatory properties of the speech motor systems between the groups. The maximum F2 deviation from the unperturbed value was 321.6±112.4 Hz and 344.6±98.4 Hz (mean ±1 SD, averaged between the Down and Up conditions) in the PWS and PFS groups, respectively, which did not differ significantly between the two groups ($p>0.4$, t-test)[32].

An additional potential confounding factor that warranted examination was differences in the timing of the perturbations. This is because differences in the duration and/or rate of increase of the perturbation made lead to differences in the compensatory responses. Therefore, we will focus on two parameters of the auditory feedback, 1) the duration of the perturbation (i.e., the length of the time from the onset of the perturbation shortly after [i] to the end slightly before $[j]_1$, see Figure 2.2.A for an example, and 2) the time lag from the onset of the perturbation to the maximum F2 perturbation (at $[u]_1$). The first measure quantifies the total duration of the perturbation; it was on average 265.5±33.9 ms and 271.9±31.6 ms (1 SD) in the PWS and PFS groups, respectively, and was not significantly different between group ($p>0.4$, t-test). The second measure captures the rate of the ramping up of the F2 perturbation from zero at the onset

[32] The magnitude of the F2 perturbation was determined by the range of F2 during the perturbation interval (i.e., during the word "owe" and the subsequent transition into the word "you"). From the fact that we found no significant difference in the magnitude of perturbation between the two groups of subjects, it can be seen that the range of F2 during the perturbed part of the sentence did not differ significantly between PWS and PFS. This finding was discrepant with previous reports of more centralized vowel space in PWS compared to controls (Klich and May 1982), but is consistent with other reports of the absence of evidence for abnormal vowel formants in the fluent running speech of PWS (e.g., Prosek et al. 1987).

and maximum at [u]1; this time lag averaged 128.8±30.0 and 132.9±26.5 ms (mean ± 1 SD) in the PWS and PFS groups, respectively, and the between group difference was not statistically significant (p>0.6, t-test).

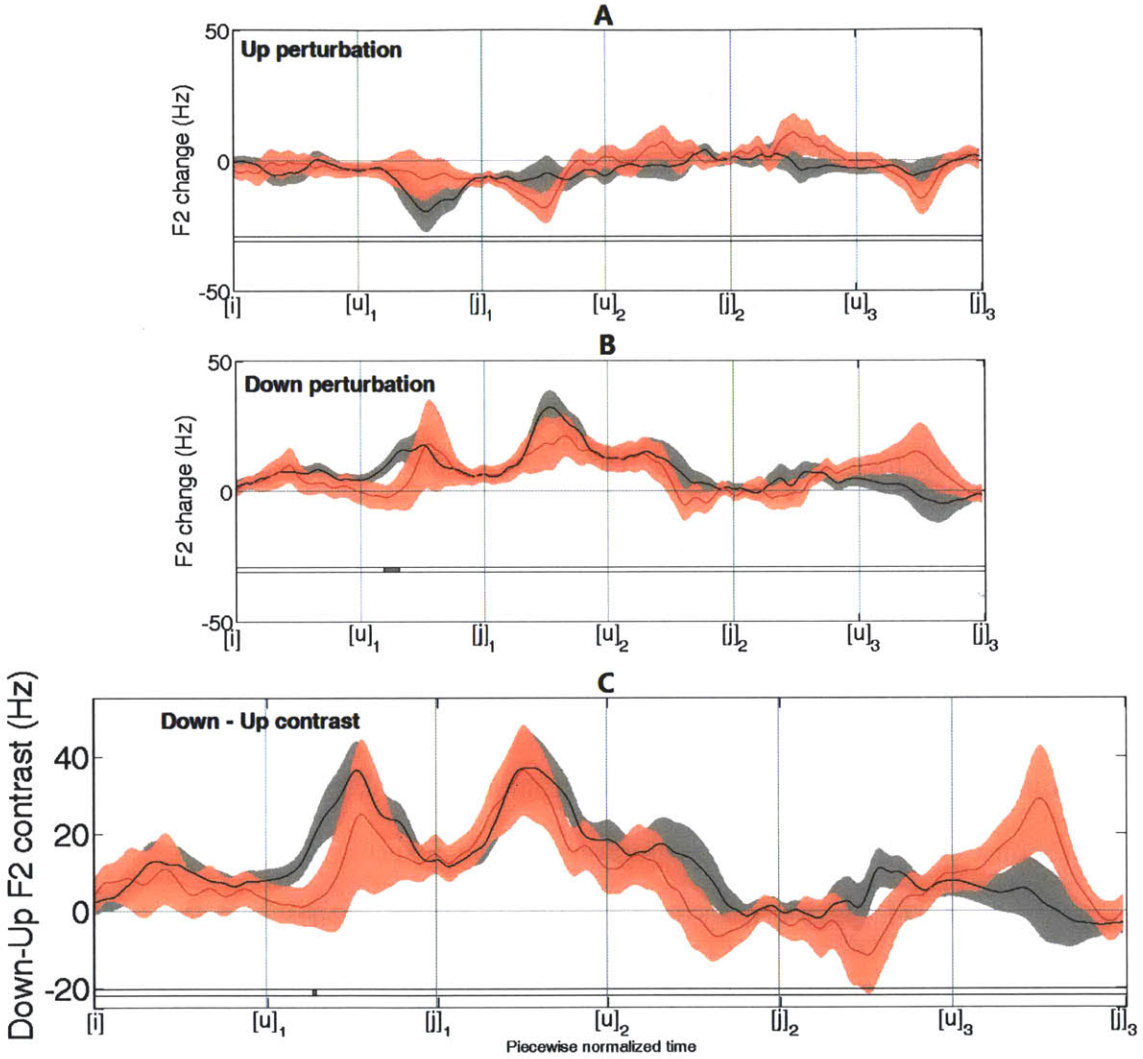


Figure 4.11. F2 compensation curves under the Up (A) and Down (B) perturbations. The F2 change curves in this figure are shown on the piecewise normalized time axis, from [i] at the end of the word "I" to $[j]_3$ at the beginning of the last syllable of the sentence ("yo").The format is the same as in Fig. 2.11. The data from the PWS and PFS groups are shown by the red and black curves, respectively. The shading around the solid curves show mean ± 1 SEM. The horizontal bar on the bottom of each panel indicate significance between the PWS and PFS groups on a frame-by-frame basis. The gray color indicates significant group difference at an uncorrected level of p<0.05 (two-tailed t-test). **C**: The contrasts between the Down and Up responses. As in Panel B, the gray area in the horizontal bar near the bottom indicates the interval in which the average response magnitude of the PWS was significantly smaller than that of the controls (same statistical threshold as in B).

As in Sect. 2.1.2.1, we separately analyze the changes in the spatial and temporal aspects of the subjects' productions under the Down and Up perturbations. To analyze the spatial changes, we use the same method as used for generating Fig. 2.11.A. The time-varying trajectories are aligned at the seven extremum landmarks ([i], $[u]_1$, $[j]_1$, $[u]_2$, $[j]_2$, $[u]_3$, $[j]_3$) and the time points between these anchor points were linearly analyzed. In the figure above, Panels A and B show the compensatory F2 changes under plotted along the piecewise normalized time axes, under the Up and Down perturbations, respectively. It can be seen that both the shapes of these F2 compensation profiles were similar between the two groups of subjects in that there were changes in the produced F2s in the directions opposite to the perturbations in both groups. However, under the Down compensation, there was a time period between the $[u]_1$ and $[j]_1$ landmarks, in which the magnitude of the compensatory F2 decrease was significantly smaller in the PWS group than in the PFS one (see the gray area in the horizontal bar near the bottom of the Panel B, $p<0.05$, uncorrected, two-tailed t-tests). This difference was observed only in the earliest part of the compensatory response. The magnitude of compensation of the PWS "caught up" with the PFS and no significant difference was found between the two groups. A similar period of significantly weaker-than-normal compensation in the PWS was found when the composite response (i.e., the Down – Up contrast) was analyzed (Fig. 4.11.C).

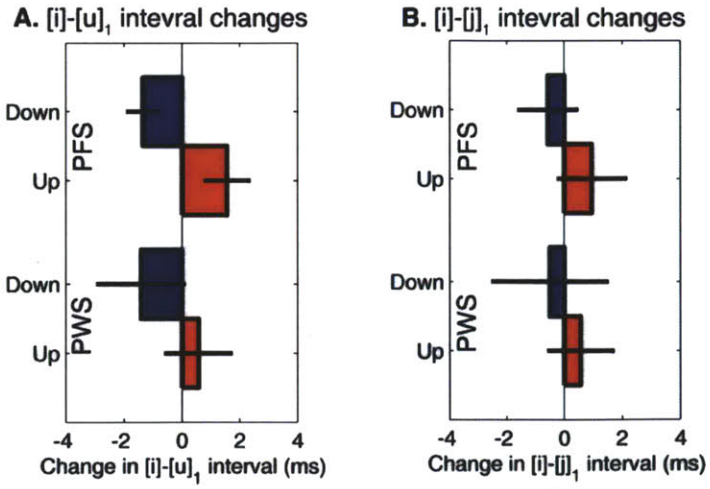


Figure 4.12. Changes in time intervals during the utterance "I owe you a yo-yo" under the Up (red) and Down (blue) perturbations from the noPert baseline. Pane A shows the changes in the [i]-$[u]_1$ interval; Panel B shows the changes in the [i]-$[j]_1$ interval. In each panel the data from the PFS group is shown on the top, whereas the data from the PWS was shown on the

bottom. The error bars show ±1 SEM around the group means. The timing change patterns under the spatial time-varying perturbations were quite similar between the groups and no significant between-group differences were found.

The above findings indicate that under the Down and Up perturbations, the PWS showed weaker F2 adjustments in an early part of the compensatory response, indicating that despite that PWS are capable of generating close-to-normal articulatory compensation, their AF-based control systems may take longer time to generate and fully implement the appropriate corrective motor commands.

In addition to the spatial adjustments, we also compared the online time adjustments exhibited by the two groups. We have previously shown in Sect. 2.1.2.2 that normal subjects showed significant changes in the [i]-[u]$_1$ interval under the Down and Up perturbations. As Fig. 4.12.A shows, PWS showed average pattern of the [i]-[u]$_1$ interval change that was similar that of the PFS controls. No significant difference was found in the Up-Down contrast of the [i]-[u]1 interval between the two groups of subjects ($p>0.6$). Similarly, the timing change patterns in the [i]-[j]1 interval was also similar and did not differ significant between the PWS and controls ($p>0.8$).

4.3.2.2. Experiment B: Responses to temporal perturbation by the people who stutterers

The PWS subjects were able to produce the stimulus utterance within the target range of level and duration in 91.8%±8.8% (±1 SD) of the trials, which was slightly lower than but did not differ significantly from the hit ratio from the PWS group: 95.1%±4.4% ($p> 0.094$, t-test).

As in the analysis of the spatial perturbation, it was necessary to examine whether the perturbation parameters differed substantially between the two groups, which if true, could have led to confounds in the between-group comparisons. To this end, we focus on two parameters, 1) the total duration of the perturbation, and 2) the amounts of timing shift in the [u]$_1$ F2 minimum in the AF due to the Accel and Decel perturbations. The total duration of the perturbation did not differ significantly between the two groups (PWS: mean±1 SD: 276.8±46.4 ms; PFS: 266.1±28.6 ms; $p>0.3$). Also, we computed the difference between the average timing shifts of the [u]$_1$ F2 trough under the AF between the Decel and Accel conditions as a measure of timing shift

amount, and found there was no significant difference this amount (PWS: mean±1 SD: 72.1±21.9 ms; PFS: 68.1±13.6 ms; $p>0.4$). Therefore it can be seen that the perturbation parameters were similar between the two groups of subjects under comparison and the concern that any between-group difference are primarily caused by differences in perturbation magnitudes or timing is minimal.

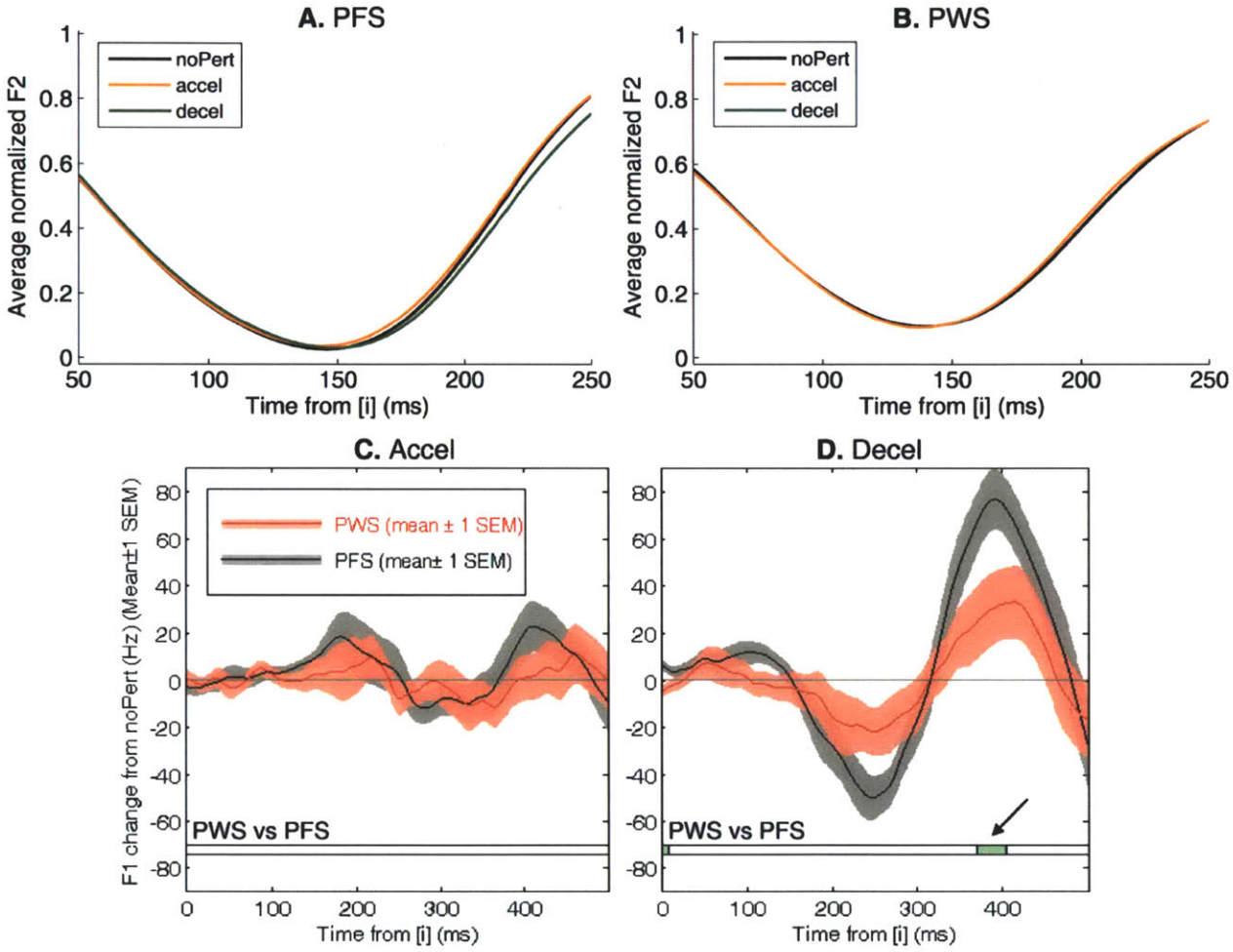


Figure 4.13. F2 trajectories produced by the PFS and PWS subjects and their changes from the no-perturbation baseline under the time-varying temporal (Accel and Decel) perturbations. **A**: the group-average F2 trajectories produced by the subjects in the PFS group under the Accel (A) and Decel (B) conditions. Before averaging, the trajectories from individual trials were all aligned at [i]: the local F2 maximum at the end of the word "I". In both panels, the group-average noPert F2 trajectory was plotted for comparison. The slight advancement in time of the produced F2 trajectory up under the Accel perturbation and the more pronounced and extensive slowing down under the Decel perturbation can both be seen. **B**: the group-average F2 trajectories from the PWS group, in the same format as Panels A and B. Through comparison with Panels A and B, It can be seen that the PWS subjects show smaller timing shifts under the Accel and Decel subjects than the controls subjects. **C:** the trajectory differences between the Accel and noPert conditions in the PFS (black) and PWS

(red) groups. **D:** the trajectory differences between the Decel and noPert conditions. The format is the same as C. The horizontal bar near the bottom of C and D shows the results of the frame-by-frame statistical comparison between the PWS and PFS groups. Color code: white: no significant difference; light cyan: significant difference at $p<0.05$ uncorrected (two-tailed t-test). The black arrow in Panel D indicates the time interval in which there is a significant difference in the trajectory changes between the two groups.

As we have shown in Sect. 2.2.2, the PFS subjects slightly accelerated the transition from the F2 maximum at [i] and the F2 minimum at $[u]_1$ in their productions under the Accel perturbation, which led to a shortening of the [i]-$[u]_1$ interval under the Accel perturbation compared to the noPert baseline. In addition, they slowed down the [i]-to-$[u]_1$ transition under the Decel perturbation, which led to a significantly lengthened [i]-$[u]_1$ interval (see Fig. 2.16.B). Figure 4.13.A below illustrates the average F2 trajectories produced by the PFS under the three different conditions. Despite the temporal smoothing in the time-based averaging, the small acceleration and deceleration of the trajectories is discernable under the Accel and Decel perturbation conditions.

By contrast, the PWS subject showed smaller timing adjustments of such type than the PFS under the perturbations, which can be seen by comparing Fig. 4.13.B with Fig. 4.13.A. In the productions of the PWS subjects, the F2 trajectories under the Accel, Decel and noPert conditions were not as separated at the $[u]_1$ trough as those from the PFS. In fact, the F2 trajectories produced by the PWS under the Accel and Decel conditions did not start to separate from the noPert baseline trajectory until about 50 ms after $[u]_1$.

The average F2 trajectory changes from the noPert baseline under the Accel and Decel perturbations are shown in Panels C and D of Fig. 4.13., respectively. The black and red curves show the data from the PFS and PWS groups, respectively. In the PFS, the changes in the F2 reflected the speeding-up and the slowing-down of the production trajectories of F2. For example, the positive deflection at about 100 ms after [i] under the Decel perturbation in the PFS group was due to the slowing down of the downward F2 sweep from [i] to $[u]_1$. By comparison, the same positive F2 deflection was nonexistent in the PWS group (red curve), which is consistent with the lack of timing adjustment in the PWS group seen before. Furthermore, the large positive peak of the PFS's F2 change curve at approximately 400 ms after the onset of the Decel

perturbation was due to the slowing down of the production between $[j]_1$ and $[u]_2$. The PWS subjects showed a positive peak at a similar time (~400 ms), but the magnitude of this peak was significantly smaller compared to the corresponding peak of the PFS group ($p<0.05$, uncorrected, two-tailed t-test, see the light green segment in the bottom bar in Fig. 4.13.D). This difference reflected smaller timing adjustment under the Decel perturbation in the PWS compared to the control subjects' timing adjustment.

This observation that PWS adjusted timing by smaller amounts based on the F2 trajectories was confirmed by the analyses applied directly on the time intervals (i.e., extracted from individual-trial trajectories). Figure 4.14. summarizes the subject-average Decel-Accel contrast of the $[i]$-$[u]_1$ interval, i.e., the differences in the $[i]$-$[u]_1$ interval between the Decel and Accel conditions. The majority of the PFS subjects showed positive values in this contrast and the average value was 4.52±0.94(SE) ms, reflecting the shortening and lengthening of this time interval under Accel and Decel in the average trend, respectively. By comparison, the values of this contrast from the PWS group were had a close-to-zero value. In fact, the average contrast was slightly negative (-0.535±1.86 ms), reflecting the virtual lack of online adjustment in the $[i]$-$[u]_1$ interval by the PWS subjects. The difference in this Decel-Accel contrast of $[i]$-$[u]_1$ interval reached statistical significance ($p=0.011$, t-test).

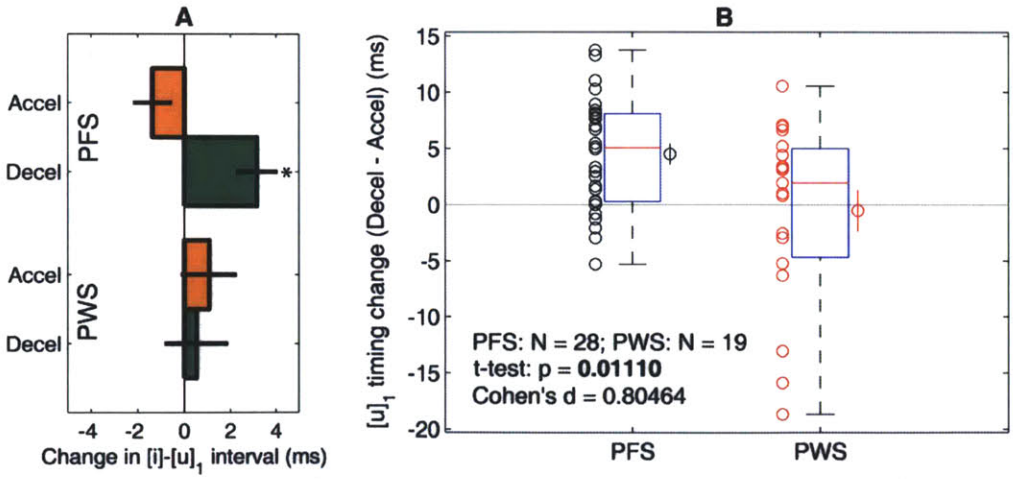


Figure 4.14. Comparison of the online timing corrections under the temporal (Accel and Decel) perturbations in the PFS and PWS groups. A. This panel has a similar format as Fig. 4.12.A and B. It compares the change in the $[i]$-$[u]_1$ interval under the Accel and Decel perturbations in the PWS and PFS groups. Notice the lack of systematic $[i]$-$[u]_1$ timing correction in the average pattern of the PWS. **B.** The left and right parts of the plot show the data from the PFS and PWS groups, respectively. The difference between the $[i]$-$[u]_1$ interval

changes under the Decel condition and that under the Accel condition (called the "Decel-Accel contrast) was computed as a composite measure of online timing adjustment in early part of the response to the AF perturbation. For the sake of comprehensiveness, each set of data are summarized in three different ways: as scatter plots (circles), Tukey's boxplot, and mean±1 SEM. The results of a t-test and a Wilcoxon rank-sum test both indicate that the PWS as a group make significantly weaker timing adjustment compared to the PFS controls ($p < 0.05$).

From the above analysis, we saw that the PWS subjects did not respond to the temporal perturbations of the AF during the earliest analyzed time interval after the onset of the perturbation ([i]-[u]$_1$ interval). In later parts of the utterance, the PWS subjects did not make timing adjustments in the same way the control subjects did, either. Figure 4.15. is a summary and comparison of the changes in the six time intervals from [i] to [j]$_3$. The upper and lower halves of this figure both have the same format as the previous Fig. 2.17., in which the filled and unfilled (open) circles indicate significant or non-significant difference from zero, respectively. The PFS group showed significant increases in all the six time intervals, from [i]-[u]$_1$ to [i]-[j]$_3$, as can be seen from the filled green circles in the upper panel (in fact, the upper half of this figure is identical to Fig. 2.17.) The average timing change pattern under the Decel condition in the PWS group (bottom panel) appears to be similar to that of the controls, but these timing changes were smaller in magnitude (especially in the early time intervals, e.g., [i]-[u]$_1$ as mentioned above) and none of them reached statistical significance. However, when the two groups were compared for each of the Decel-Accel contrast of six individual intervals, only the first one, the [i]-[u]$_1$ interval reached statistical significance. Therefore it can be inferred that the primary deficit of the PWS in the online timing adjustment is that they are slightly slower in initiating the timing correction. This pattern of smaller-than-normal adjustment magnitude in early part of the compensation accompanied by gradual catching-up of the normal response pattern is reminiscent of the between-group difference we observed in the spatial perturbation experiment (see the previous section). The [i]-[u]$_1$ interval was the earliest extremum-based time interval following the onset of the perturbation. For this early response, the PWS showed smaller-than-normal adjustment compared to controls. But for later ones, the between-group

difference in timing corrections was smaller. Therefore it can be seen that the PWS group gradually "caught up" with the control group in the average timing correction pattern during later part of the utterance.

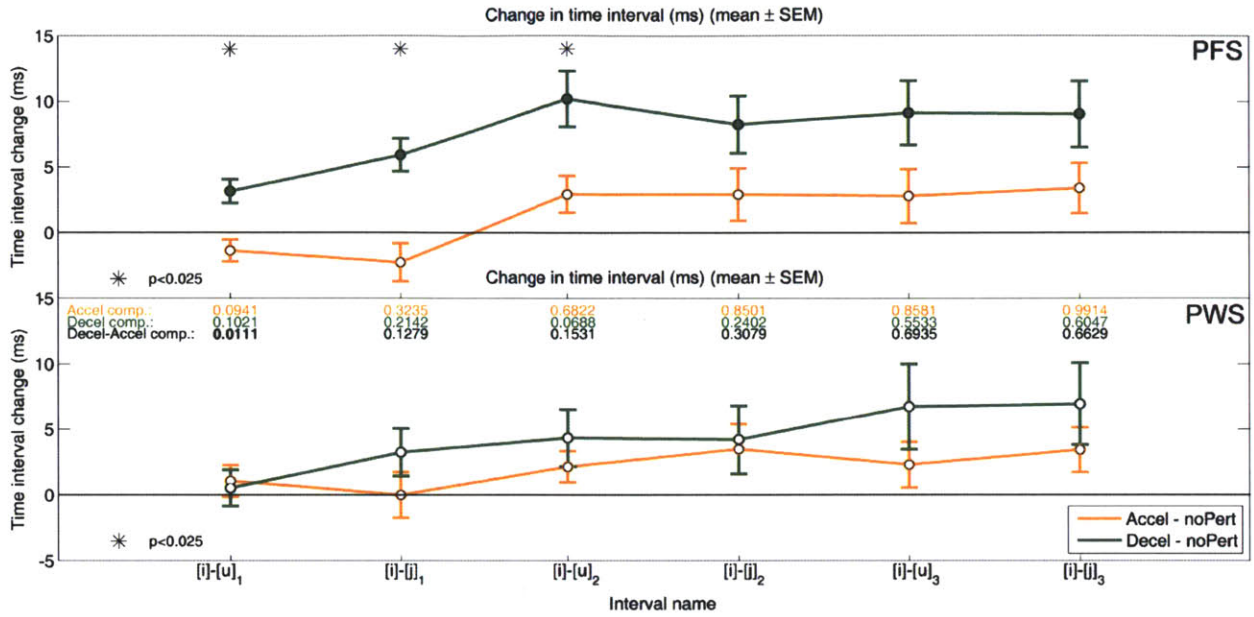


Figure 4.15. Changes in time intervals during the utterance "I owe you a yo-yo" under the Accel (orange) and Decel (Green) perturbations from the noPert baseline. The top and bottom panels show the data from the PFS (control) and PWS groups, respectively. Filled circles indicate significant difference from the zero baseline ($p<0.05$, Tukey's HSD) following RM-ANOVA. Asterisks on the top indicate significant difference between the Down and Up conditions for a time interval (within each subject group). The numbers on the top of the bttom panel are the p-values from the t-test comparisons between the two groups. The three rows show results from the between-group comparisons for the three contrasts: Down-noPert (blue), Up-noPert (red), and Up-Down (black).

The sqDIVA-T model described in Chapter 3 was fit to the data from the PWS group, and the values of the fit parameters, including the latency D, the DIVA-style feedback control weight w_{FB}, the timing adjustment coefficient w_T, and the between-/within-syllable control ratio were compared between the groups. The only significant between-group difference we found was a longer response latency in the PWS compared to the controls ($p=0.018$, two-tailed t-test, see Fig. 4.16), corroborating two observations: 1) PWS were able to make timing adjustments in their articulation in response to the temporal perturbations in their AF, and 2) they were significantly slower than fluent controls in initiating such compensatory motor corrections.

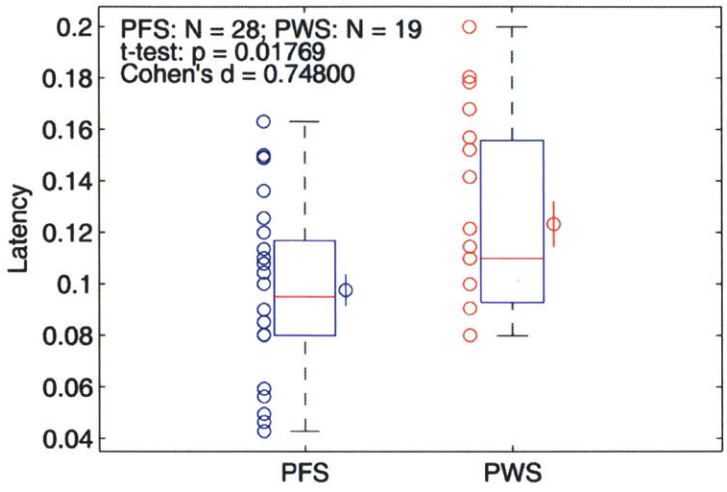


Figure 4.16. Comparing the latency of response fitted with the sqDIVA-T model to the timing-perturbation responses from the control (left) and PWS (right) subjects. The individual circles show latencies fitted on the individual subjects. The circle with the error bars show ±1 SEM.

4.4. Discussion

4.4.1. Sensorimotor integration in speech and non-speech movements of stutterers

Summarizing the findings from Experiments I and II, we conclude that stuttering is associated with a weaker-than-normal online compensatory responses to perturbations of AF. Significant differences in the amount of the compensation were found under both the F1 perturbation during a quasi-static vowel (Experiment A) and under the time-varying F2 perturbations during a multisyllabic utterance (Experiments B and C). These findings constitute evidence that the deficit is not restricted to a single aspect of speech motor control, but applies generally to both static gestures and gestural transitions, as well as to the control of spatial and temporal parameters of articulatory movements.

With regard to the mechanisms of this between-group difference, there are (at least) three possible interpretations of what specific anomaly of the speech motor system may lead to these abnormally weak auditory-motor compensatory responses.

1) First, it may reflect an anomaly in the way the auditory system processes information from the AF (i.e., primarily a perceptual deficit)

2) Second, it may be due to a limited flexibility of the speech motor system in stutterers in implementing corrective online updates to motor programs (i.e., a primarily motor deficit).

3) Third, it may be indicative of certain deficits in the neural mechanism that translate error information from AF into proper corrective motor commands, i.e., deficits in the function of an auditory-motor inverse internal models (Perkell et al. 1997; Max et al. 2004; Ventura et al. 2009; Hickok et al. 2011).

With regard to the first possibility, prior research on auditory perception in stutterers has yielded mixed results (reviewed in Chapter 6 of Bloodstein and Ratner 2008) that together are challenging to interpret. However, the fact that in the perceptual discrimination test described in Sect. 4.2.2.3 we failed to find between-group differences in the auditory acuity to vowel F1 changes provides evidence against the possibility of a purely perceptual deficit. In this perceptual task, the auditory systems of PWS seem to perform similarly to controls in detecting the type of formant shifts used in the AF perturbation experiment. Therefore, if the weaker-than-normal online auditory-motor compensation in the PWS is indeed attributable to certain deficits that resides purely within the auditory system, it would be a deficit that is manifested only during its interaction with the motor system.

The second possibility is even more tenuous, given the previous observation that measures of the trial-to-trial variability of the speech acoustic parameters are in fact greater in PWS than in PFS (e.g., Kleinow and Smith 2000; Cai et al. 2011). In the static-vowel perturbation experiment, we observed greater variability of F1 in the PWS group than in the control group under all three conditions (see Fig. 4.17. below). In addition, PWS also showed greater-than-normal trial-to-trial variability of the timing of the syllables when producing the utterance "I owe you a yo-yo" in Experiment B (not shown), a result consistent with Wieneke et al. (2001). Given that prior

studies have shown greater trial-to-trial spatiotemporal variability of articulation in PWS than in PFS (Kleinow and Smith 2000), it can be concluded that PWS are indeed *not* more inflexible than PFS in either the spatial or temporal domains of articulatory movements, hence the smaller-than-normal compensation observed in the current study cannot be attributed to a purely motor deficit, either.

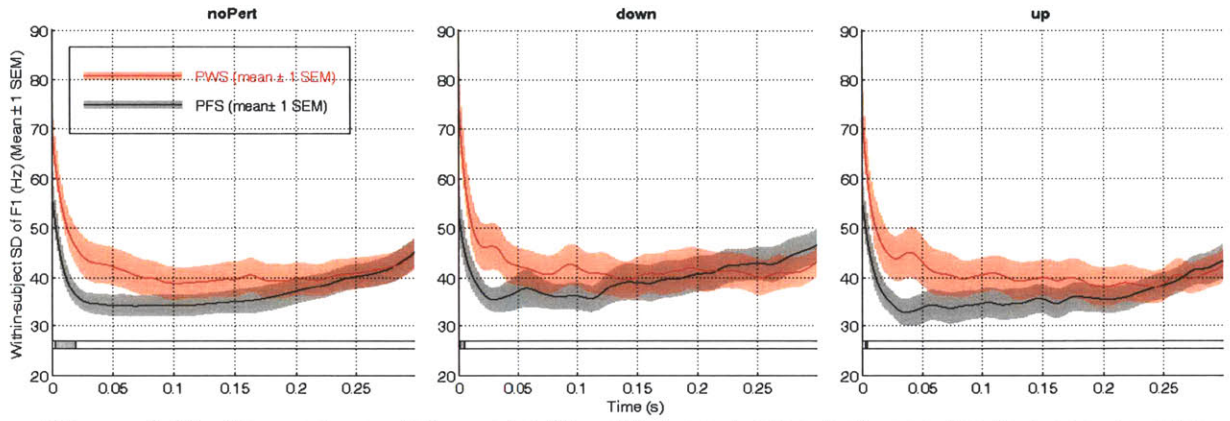


Figure 4.17. Comparison of the variability of F1 production between the two groups. The average trial-by-trial standard deviation (SD) of F1 in the subjects' productions computed on a time-frame-by-time-frame basis, under the three different conditions: noPert (left), Down perturbation (middle), and Up perturbation (right). The red and black curves show the data from the PWS and control groups, respectively. The shadings around the curves show ±1 SEM. The bar near the bottom of each panel shows the significance of between-group t-test comparisons (two-tailed) between the PWS and controls. The gray areas indicate time periods in which the between-group difference reached the level of $p<0.05$ (uncorrected). Within the first 25 ms following the vowel onset, the PWS showed significantly higher trial-to-trial variation than controls. Overall, there is no evidence that the speech motor control is more rigid (less flexible) in PWS than in controls, as the second possibility (pure motor deficit) would suggested (see text for details).

In the light of the above discussion, the third interpretation mentioned above, namely a defective auditory-to-motor transformation mechanism, seems to be the most parsimonious and compelling possibility. According to this interpretation, neither the encoding of auditory information nor the capacity to implement corrective motor commands are defective in stuttering, but the neural processes that translate the sensory error information into the appropriate corrective motor command is dysfunctional or noisy. This idea of an impaired / inadequate internal auditory-motor transformation is not an original one. A couple of prior papers have

articulated the basis of this idea (Max et al. 2004; Hickok et al. 2011). But to our knowledge, this is the first systematic, unequivocal empirical evidence to support it.

Kalveram (1987) developed a model for abnormal use of AF by the speech motor system for timing control in PWS. His model was supported by the finding that PWS show atypical responses to DAF (Kalveram and Jancke 1989). The findings of the current study are largely consistent with the premise that interaction between AF and speech motor timing control are abnormal are different, but contradicts with the findings of Kalveram and colleagues in that our results tend to show slower and weaker online updating of timing in PWS than in normal speakers, whereas the findings of Kalveram and colleagues tended to show an abnormal timing of AF-based control. In fact, Kalveram and colleagues showed that PWS' speech timing was more affected by DAF than normal speakers' speech was (Kalveram and Jancke 1989). This discrepancy of results may be related to the different types of auditory perturbations used: whereas the perturbation used in the current study were well-controlled, phoneme- and formant-specific, and unnoticed by most of the participants, the DAF used by Kalveram and Jancke (1989) was certainly less subliminal and more coarse, and hence of doubtful generalizability to speech under the normal AF condition. The discrepancy may also have to do with the different phoneme types used in the two studies: whereas we used uttereances consisting of vowels and semivowels in this study, Kalveram and Jancke's stimulus utterance contained stop consonants. There may be intrinsic differences in the control mechanisms underlying vowel-like and stop consonants, which need to be elucidated in future studies.

The most recent theorization on this hypothesis by Hickok et al. (2011) is based upon the "State Feedback Control" (SFC) model (Ventura et al. 2009). According to the hypothesis proposed by Hickok and colleagues, an auditory-motor translation module in the speech motor system is responsible for the bidirectional information conversion (see Sect. 1.3.2 and Fig. 1.3). On one hand, it is a part of the pathway that translates efference copies of speech motor commands into predicted AF. On the other hand, it is also responsible for generating corrective motor commands based on mismatch between the auditory prediction and the true AF. Hickok

and colleagues hypothesized that stuttering results from noisiness in this module. The implication of this hypothesized noisiness has been discussed in Sect. 1.3.2. In the context of AF perturbation, this model of stuttering may predict either smaller-than-normal or noisier-than-normal compensatory adjustments. Due to the current lack of quantitative (computational) details, it is unclear which of the two will be the prediction of this SFC-based model. Nonetheless, it is conceivable that the weaker-than-normal auditory-motor adjustment observed in the current study results from either (a) inaccurate internal prediction of the auditory consequence, or (b) inaccurate compensatory motor commands based on the error (mismatch), or (c) both. All of these three schemes are largely consistent with the SFC-based hypothesis of Hickok et al. (2011).

In the context of the DIVA model, which is different from SFC, interpretations of the experimental results are slightly different. In DIVA, AF are compared to prelearned auditory, rather than to internally generated predictions of AF, to give rise to auditory errors. Therefore at least in the context of the production of single phoneme or single syllables, the possibility (a) mentioned above is incompatible with DIVA. However, possibility (b) is still largely consistent with DIVA. According to recent versions of the DIVA model, the error control map is situated in the right ventral premotor cortex (vPMC, Golfinopoulos et al. 2009). Therefore if there is indeed a deficit in the neural mechanism of computation of error-based corrective motor commands, DIVA would predict that such a deficit is a function of certain anomalies either in the right vPMC or in the connections between vPMC and other parts of the brain, such as right planum temporal (PT), which was also involved in the AF-based online speech motor compensation (Tourville et al. 2008).

However, it needs to be pointed out that DIVA, in its current form, deals mainly with short, monosyllabic utterances (cf. units stored in the *mental syllabary*, Levelt 1994). As such, it is still incapable of simulating the construction longer, multisyllabic utterances from these shorter units of articulation. Chapter 3 of this dissertation was aimed at filling in this gap. Auditory goals (templates) of timing patterns presumably do not exist for multisyllabic utterances, which are enormous in number due to the combinatorial nature of such utterances. Logically, this implies

that on the inter-syllabic level, the actual AF cannot be compared with a prelearned and pre-stored auditory target. Instead, AF has to be compared with a prediction that is generated based on the motor plan or program. For example, with the type of stimulus utterance used in Experiments A and B, the timing information in AF with regard to the onset and offset timing of syllables must be compared with an predicted AF that is generated on the fly by an internal model using the preplanned timing score (see Fig. 3.1.B). In this sense, both the SFC and DIVA model need to rely on internal forward models to generate predictions of AF during multisyllabic speech utterances.

In light of a few prior studies, the above-discussed deficit in the internal transformation from sensory to the motor domain may be not restricted to the AF, but instead be generalizable to the transformations between other sensory modalities and the motor domain. By using an unanticipated mechanical force load to the lower lip during the production of a bilabial stop consonant [p], Caruso and colleagues (1987) demonstrated that three PWS subjects showed significantly reduced compensations in the EMG activities of the lower lip and significantly longer response latencies compared to three control subjects. The small sample sizes employed by Caruso et al. (1987) left much to be desired, but the fact that statistical significance was reached even with such a small sample size may be indicative of the large effect size of this between-group difference. In another related study, Bauer et al. (1997) reported the preliminary finding that two severe stutterers (out of 10 PWS used in that study) failed to compensate for unexpected mechanical perturbation during the production of "sasasar". Qualitative similarity between these findings and the finding of weakened vowel formant compensation and timing adjustment in the current study is intriguing and seems to be hinting at a generalized sensory-motor translation deficit. It is possible that certain brain areas that are involved in the integration of modality-independent sensory information with ongoing motor control is defective in stutterers and this defect may be the common underlying cause for the findings both in the current study and in Caruso et al. (1987).

Furthermore, it is possible that this defective integration of auditory information with ongoing motor control is not restricted to speech movements, but instead exists in broader, general types of movements. Results from several previous studies are consistent with this hypothesis. These studies were based on various types of non-speech motor task that requires the concurrent monitoring and utilization of auditory information. In the study by Neilson and Neilson (1979), stuttering and non-stuttering subjects were required to make jaw or hand movements to match the pitch of a simultaneously presented tone. It was found that PWS showed significantly larger phase lags in this real-time auditory-motor following response than fluent controls. Interestingly, no difference in following response was found if the target to follow was presented visually. In another study by Nudelman et al. (1992), subjects were asked to match the pitch of their humming to the pitch of a simultaneously presented tone, and the group of four PWS showed greater phase lags in this pitch-matching humming task than controls. As mentioned in Section 1.3.1, Loucks and De Nil (2006) found that PWS perform less accurately and more variably than controls on a task that required subjects to move the jaw to a visually presented target position. These above-mentioned tasks may seem contrived and not so relevant to speech production *per se*, but they had an important commonality with speech motor control, namely the online translation of auditory information into an appropriate matching motor action for controlling the end-effectors (hand or jaw), and/or online prediction of the consequences of motor programs that are about to be issued and compared these prediction with targets in the sensory domain, which are both functions subserved by internal IMs. Therefore these non-speech sensorimotor findings seem to hint that the weakened auditory-motor adjustments seen in PWS reflect deficits in the IMs of the speech motor system, which may be a part of deficits in the motor system in general.

4.4.2. Relations to Core Behaviors of Stuttering

The current study focuses on the fluent speech of stutterers. Only the trials containing fluent productions of the stimulus utterances were included in the data analysis. Despite this, significant differences were found between the PWS and PFS groups, which indicates that the sensorimotor deficits reflected by the under-compensation persists even in the absence of overt dysfluency behaviors. However, for any hypothesis to be a viable etiological theory of stuttering, or at least to be an important part of it, the hypothesis ought to be capable of explaining how the core stuttering behaviors (e.g., dysfluencies such as sound repetitions, prolongations, and blocks) arise. Based on the DIVA model, Civier et al. (2010) proposed that dysfluencies in stuttering may result from attempts to halt and reset the production when there are the excessively large auditory errors, which occur in stutterers because of an abnormal over-reliance on auditory feedback for ongoing speech motor control. The study by Civier et al. was unique in the stuttering field in that they showed model-based quantitative simulation results to support their hypothesis. Unfortunately, the findings of the current study do not seem to support the over-reliance hypothesis (Civier et al. 2010; Max et al. 2004), at least not in any straightforward way, as smaller-than-normal, rather than stronger-than-normal compensatory responses were observed.

According to the SFC-based hypothesis of Hickok et a. (2011), overt dysfluency events in stuttering arise as a consequence of either 1) noisy predictions generated by the forward IM, which causes the speech motor system to detect false auditory errors and attempt to compensate for them when the AF actually contains no error, or 2) inaccurately computed speech motor commands based on mismatches between auditory targets and sensory state information. The above two causes may coexist and together form a vicious cycle, which may explain the temporally extended halting of speech flow and inability to transition to the next sound in sequence in prolongation-, repetition- and block-type dysfluencies. Despite the fact that the SFC model is conceptually different from DIVA in many as aspects (reviewed in Sect. 1.2.2), there is nothing in our current data that argues against the above SFC-based hypothesis.

The abnormally weak responses to perturbations of AF in PWS found in the current study provided partial empirical support to the first possibility mentioned above, i.e., the hypothesis regarding impaired inverse IMs and the erroneous error signals generated by them. However, because the current study was not focused on moments of dysfluency, it could not generate evidence for or against the mechanistic pathways proposed by the SFC model that is hypothesized to lead from the impaired IMs to the dysfluent speech. Direct empirical test of these hypothesized mechanisms will be challenging, since it will be difficult to directly measure the motor commands generated by the IMs. However, it is possible to perform indirect tests, e.g., by using non-speech oromotor tasks. In fact, previous findings such as less accurate jaw target reaching in PWS (Loucks and de Nil 2006) are largely consistent with the SFC-based hypotheses.

Another possible way in which defective sensorimotor IMs may lead to dysfluencies is through corrupted internal feedback signals for triggering ensuing syllables in a multisyllabic speech utterance. The delaying effect of the Decel perturbation we found in Experiment B of the current study and the delaying effect of an unanticipated sudden deprivation of AF (Perkell et al. 2007) are both consistent with a role of AF in the triggering of the end of an articulatory gesture and the initiation of the following one. It needs to be clarified that we are not arguing for a purely auditory-motor chain, in which the signal of completion of a syllable is necessary for the initiation of the following one. Instead, we are proposing a model in which not only auditory feedback, but also the efference copies of motor commands, possibly as well as the auditory predictions made by the forward IMs, play a role in the triggering ensuing syllables. This idea is schematically shown in Fig. 4.18.

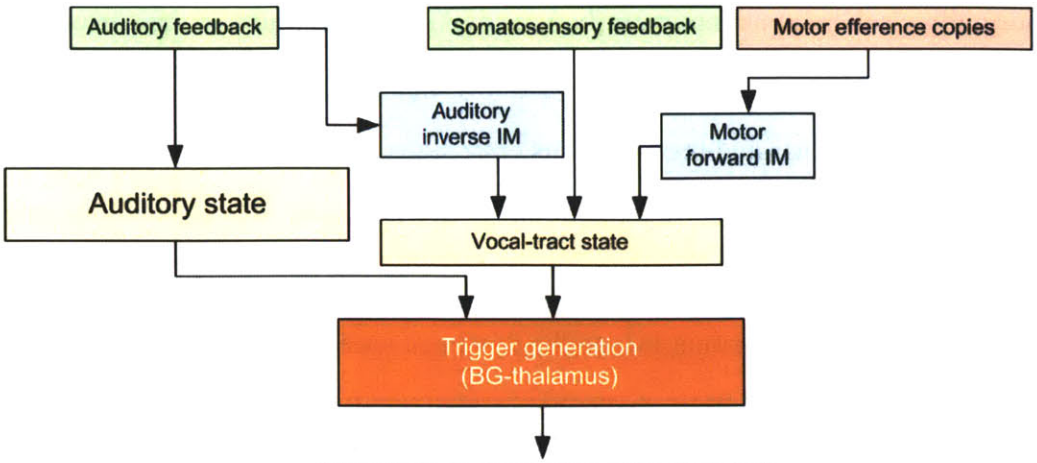


Figure 4.18. A schematic diagram illustrating our hypothesis regarding the mechanisms underlying the transitions between different syllables in a multisyllabic speech utterance. The BG-thalamic loop utilizes two types of state information to compute the appropriate timing of syllable onsets and offsets: auditory state and vocal-tract state. The computation of the vocal-tract state involves the weighting and averaging of the information from three channels: 1) results of the forward modeling based on AF, 2) direct vocal-tract information from somatosensory feedback, and 3) forward modeling results based on motor efference copies. This model has the merit of being compatible with the effects of the Decel perturbation on syllable timing and with the findings of Perkell et al. (2007) based on sudden switching of CIs. In addition, this model is also consistent with the fact that fluent multisyllabic articulation is possible even under the masking of AF and blocking of somatosensory feedback: the BG-thalamic loop is flexible and can utilize only the available state information to generate the timing triggers according to the circumstances.

In the conceptual framework schematized in Fig. 4.18, the defective auditory-motor translation may lead to dysfluency events in a number of possible ways. First, it is possible that the impaired WM connections (Sommer et al. 2002; Chang et al. 2008; Watkins et al. 2008) between the auditory cortex and the basal ganglia leads to corrupted *auditory state-dependent* trigger signals, which occasionally results in moments of dysfluency. Second, noisy or inaccurate predictions of AF generated by an inadequate IM may lead to occasional failures of the *vocal tract state-dependent* triggering. This model is actually compatible with the fluency enhancing effects of a number of manipulations. Masking noise causes the system to temporarily disengage the reliance on AF, therefore freeing the system of occasional failures due to unreliable AF-based triggers. A similar explanation can be offered for DAF, since the unnatural timing of AF in DAF may temporally render the AF signal unusable as well. It is noteworthy that several previous studies have shown that the fluency enhancing effect of altered AF gradually wears off

with extensive period of using DAF devices (Armson and Stuart 1998). It is possible that the system adapts to the unnaturally long feedback delay and recovers its utilization of AF with time. As for the strong fluency enhancing effects of singing, choral reading, and rhythmic pacing of speech (e.g., Hanna and Morris 1977; Stager et al. 1997), the sensorimotor-based triggering mechanism schematized in Fig. 4.18 may be temporarily bypassed and replaced by a central clock, which is not affected by the defects in either the AF or the IMs.

It will be possible to perform an indirect test of an important premise of the failure-of-trigger hypotheses mentioned above in future studies. For example, PWS may be especially susceptible to the corruption of auditory state-dependent triggers during the production of multisyllabic speech. By using unexpected, sudden-onset manipulations of AF (e.g., formant shifts or delaying) and observing whether such manipulations increase the chance of dysfluency in PWS' speech, we can perform a test of the hypothesis that dysfluencies in stuttering (or at least a subset of them) are due to failures of generating proper triggers for ensuing syllables due to erroneous auditory state information. However, in such experiments, care needs to be taken to ensure that

1) Speech utterances are not produced in a repetitive fashion (as in the current study), so as to minimize the likelihood that a complex multisyllabic utterance is overlearned or chunked into a single unit, which may reduce the role of sensory state-dependent trigger and hence obscure the effects of the feedback manipulation;
2) The perturbations are introduced with an unexpected timing, so as to reduce the possibility that the subjects may anticipate the arrival of the perturbation and consciously or subconsciously employ strategies to cope with them.

To our knowledge, no perturbation experiments that satisfy the two above-mentioned criteria have been conducted before: existing studies based on randomized perturbations of AF (including the current one) didn't meet the first (non-repetitiveness) criterion; previous studies based on DAF and noise masking didn't satisfy the second (unexpectedness) criterion. We think conducting a new study that meet these two criteria will generate important and useful insights

into the sensorimotor and neural mechanisms of multisyllabic articulation in normal speech and stuttering. Oral reading of previously unseen paragraphs or topic-constrained spontaneous speech should be a speaking task that serves this purpose well. Unexpected formant manipulations or focal delays (as those used in the current study) can be introduced during such oral reading tasks to test whether unanticipated corruption of sensory states lead to breakdown of fluency.

4.4.3. Possible neural correlates of the impaired sensorimotor integration

Before discussing the possible neural correlates underlying the auditory-motor under-compensation in PWS, it is necessary to first briefly review the current status of our knowledge of the neural underpinning of auditory-motor interaction in normal speech production. By using real-time AF perturbation during fMRI, several previous studies have indicated the involvement of bilateral pSTG, PT, as well as right lateralized cortical areas and inferior posterior cerebellum, in the online control of speech movements based on AF (Toyomura et al. 2007; Tourville et al. 2008). MEG studies have shown that during speech production, the magnitude of the M100 response, of which the source is localized to the superior temporal lobe, is suppressed relative to passive listening to recordings of self-produced speech. But when the AF of the vocalization is manipulated and rendered different from the natural feedback in certain acoustic parameters (e.g., shifted pitch), the magnitude of this suppression will be reduced (Paus et al. 1996; Curio et al. 2000; Houde et al. 2002; Heinks-Maldonado et al. 2006; Christoffel et al. 2007). It is noteworthy that the suppression of auditory cortical responses to self-produced sounds (referred to as motor-induced suppression, or MIS) and the weakening of this suppression by perturbation of AF are present not only in humans, but also in vocalizing primates, as indicated by extracellularly recorded single-unit spiking activities (Eliades and Wang 2005, 2008). These findings provide strong evidence for a central role of posterior superior temporal regions, such pSTG and PT, in the online processing of AF for integration with motor control.

Unfortunately, most of the previous studies on auditory-motor integration in speech have been using quasi-static articulatory or phonatory tasks; very few prior studies have explicitly

examined the neural substrates of auditory-motor interaction during multisyllabic speech. Therefore we know very little about how the brain performs online AF-based control of time-varying articulation. Future studies are required to attain important knowledge in this area.

There is evidence that the MIS of auditory cortical responses is abnormal in adults who stutter. Beal et al. (2010) observed that adults who stutter show shorter latencies (by ~25 ms) of the M100 response to self-generated vowels in the right hemisphere compared to the left, but there was no such bilateral latency asymmetry in fluent adults. In addition, adults who stutter showed longer latencies in both hemispheres in the passive listening (both listening to vowels and listening to words) task compared to controls. The same group of researchers (Beal et al. 2011) also reported that children who stutter show longer M50 latencies than children who do not stutter, when collapsed across a vowel listening and a vowel production (suppression) task. These latency abnormalities, including abnormally long latencies and left-right asynchrony during speech motor performance, may be related to the under-utilization of AF for online speech motor compensation observed in the current study.

As a technique with high spatial resolution compared to MEG, MRI can better localize the neuroanatomical foci of the anomalies. By using structural MRI, Foundas and colleagues (2004) have identified a macroscopic structural anomaly of the auditory cortical regions in the superior temporal plane and their findings also shed light upon the possible functional implication of the anatomical anomaly. Based on manual demarcation in MRI images of the planum temporal (PT) in both hemispheres, Foundas et al. (2004) observed that in a group of PWS subjects, atypical asymmetry (right > left) of the PT volumes is correlated with higher level of dysfluency under normal AF and enhanced fluency in responses to DAF. The PWS with typical (left > right) asymmetry of PT volume were more fluent than the PWS with atypical PT asymmetry and did not show significantly improved speech fluency under DAF. With the caveat that the low dysfluency level of the PWS with typical PT asymmetry might have caused a "floor" effect and masked the fluency enhancing effect of the DAF, these findings indicate that the morphological abnormalities in bilateral PT may have close relations to abnormal auditory-motor interaction in stuttering.

Functional neuroimaging studies, mainly using fMRI and PET techniques, have repeatedly reported that PWS show atypical activation in the posterior superior temporal areas during

speech production[33]. These abnormal functional activations in auditory cortex are summarized in Table 4.1. As can be seen from this table, the most common finding across these studies is an underactivation of the pSTG, especially the left pSTG (BA22).

Table 4.1. A survey of abnormal auditory cortical activation in the superior temporal regions during speech production in people who stutter.

Citation and imaging modality	**Task**	**Main finding related to auditory cortical activation**
Fox et al. (1996): $^{15}O_2H$ PET	Oral reading of paragraphs: solo and chorus reading.	Left superior temporal activation was much weaker in PWS than in controls. In addition, PWS show weaker activation in what the authors referred to ask left inferior lateral premotor (ILPrM, BA6/BA44) area than controls.
Braun et al. (1997): $^{15}O_2H$ PET	Narrative and sentence completing tasks	Unlike fluent controls, PWS failed to activate the left pSTG (BA22) during the speech task. In addition, in the stuttering group, there was a negative correlation between the severity of stuttering and rCBF in the right pSTG.
Stager et al. (2003): $^{15}O_2H$ PET	Fluency-inducing conditions of speech production: rhythmic pacing (92 beats/min) and singing a nursery rhythm.	During the fluency-inducing speaking conditions, PWS show greater rCBF in the left pSTG (BA22) than controls.
Fox et al. (2000) $^{15}O_2H$ PET	Oral reading of paragraphs (solo and chorus)	Negative correlation between rate of stuttering and rCBF found in bilateral pSTG.
De Nil et al. (2000): ^{15}O PET	Oral reading of three-syllable words	Weaker activation in the left STG in PWS than in controls. Weaker activation in the left IFG in PWS than in controls.
Watkins et al. (2008): fMRI	Oral reading of sentences, under delayed and frequency-shifted AF. Sparse sampling paradigm.	No direct auditory cortex-related findings, perhaps as a consequence of the unnatural AF condition they used in the fMRI task.
Chang et al. (2011): fMRI	Oral reading of nonsensical, disyllabic words. Sparse sampling.	Weaker left BA44 to right STG functional connectivity during the task in PWS than in controls.

[33] Unlike MEG and EEG, fMRI and PET do not measure signals that are directly caused by neuronal activities. Both fMRI and PET measure quantities that change with local blood flow and blood volume, which in changes in response to neural activity changes, albeit on a much slower temporal scale than the electrical or magnetic signals measured by EEG and MEG (Huettel et al. 2008). The relation between the hemodynamic response and the underlying neural response is a complex one. But currently most researchers accept that the BOLD and rCBF signals measured by fMRI and PET are more correlated with the strength of synaptic events (hence presynaptic neuronal firing) than with action-potential firing activities of the neurons (Logothetis et al. 2001)

Apart from simple region-by-region activation analyses, researchers have recently begun to devote attention towards the functional connectivity among different brain areas during speech and how these connectives differ between PWS and fluent speakers. Chang et al. (2011) use psychophysical interaction (PPI) to compare the changes in the functional connectivity among different brain areas between an overt speech production (oral reading of nonsensical disyllabic words) and silent conditions. By contrasting the PPI results from the PWS and controls, Chang and colleagues revealed that the speech-related increased in functional connectivity between left BA44 (the seed region) and right STG (among other target brain areas) was significantly weaker in PWS than in controls. Lu et al. (2009) used probabilistic ICA (PICA) and structural equation modeling (SEM) to study the differences between PFS and normal speakers in a large-scale functional connectivity patterns during speech production. The results of the PICA analysis differed vastly between the stutterers and nonstutterers. Whereas in the control group, the bilateral pSTG was included into the first independent component (with negative correlation with the time course of the component), the same brain area did not appear in the independent component extracted from the stuttering group. In addition, results of the SEM analysis indicated that the connection between the left pSTG and the left putamen/thalamic regions was significantly different in PWS and PFS.

These functional activation and connectivity abnormalities in the temporal auditory areas during speech may be closely related to the behavioral under-compensations observed in the current study. However, these abnormal neural activities also beg the question of what are their underlying causes.

A definite answer to this question remains elusive at this point, given our still-primitive understanding of the speech motor system, even in healthy normal individuals. However, several possible answers arise in the light of other previous findings based on structural MRI, in particular diffusion tensor imaging (DTI).

First, it is possible that the abnormal functional activity in superior temporal areas during speech is a result of defective axonal connection with other speech-related areas. Chang and

colleagues (2011) used probabilistic tractography analysis (Behrens et al. 2007) of DTI data to examine the WM tracts that connect the left BA44 with other regions of the brain. Comparing the tract densities in PWS and PFS, they found that the tract density emitting from the left BA44 to other left-hemisphere area was substantially less in the PWS than in controls. In particular, the tracts that led from left BA44 to the left superior temporal regions were significantly less pronounce in PWS than in controls.

With regard to left IFG pars opercularis, which corresponds to cytoarchitectonically defined area BA44 and often considered to be part of Broca's area (Amunts et al. 1999), a few studies have shown structural abnormalities of this key area of the speech motor system in PWS. For example, Kell et al. (2009) found reduced gray matter volume in the IFG (BA44) in adults who stutter compared controls by using Voxel-based Morphometry (VBM, Ashburner 2000). In addition, they found that the gray-matter (GM) volume of the same brain area was lower-than-normal even in a group of subjects who had recovered spontaneously from developmental stuttering. Moreover, in the PDS group, there was a significant negative correlation between the GM volume in left BA44 and stuttering severity. Interestingly, in children who stutter (CWS), the same brain area shows reduced GM volume compared to age-matched fluent controls (also based on VBM, Chang et al. 2008). This convergent pattern of observations across studies indicates that the structural anomalies in the GM of left IFG is likely to be closely related to the primary etiology of this disorder.

This reduced functional and structural connectivity between the left IFG and auditory areas reviewed above may be related to the weaker-than-normal compensations to AF perturbations observed in the current study. According to the DIVA model, BA44 is the one of the primary loci of speech sound map (SSM)[34], which supplies sensory areas with sensory targets during speech. Based on this model, a disconnection between BA44 and the superior temporal areas (pSTG and PT) will result in the failure of sensory prediction, which can in turn lead to under-compensation to sensory perturbation such as the findings of the current study.

[34] The other primary loci of the SSM is the left ventral premotor cortex (vPMC).

Finally, the finding of weaker timing adjustment in PWS deserves some additional discussion. As mentioned above, the current knowledge about neural substrates underlying the sub-second online timing control is lacking. However, there is evidence that perception and motor production of sub-second time intervals share a common mechanism (Keele et al. 1985; Ivry and Hazeltine 1995). People who are more accurate at perceiving differences in sub-second time intervals tend to be more accurate at performing tasks that require the production of sub-second time intervals (e.g., with finger tapping). Several previous studies have compared timing perception abilities between PWS and fluent controls (Herndon 1966; Ringel and Minifie 1966; Barasch et al. 2000; Ezrati-Vinacour and Levin 2001). A wide range of time intervals, from sub-second to as long as 30 seconds, were examined in these studies. Perhaps the study most relevant to the current one is Herndon (1966), who administered the Seashore Measures of Musical Talents on a group of PWS and controls. This test involving a two-interval two-alternative forced choice task, in which the subject judged which of a pair of tones, played sequentially, is longer. The two tones in a pair were based on a standard duration of 800 ms but differed by an amount in the range of 50 to 300 ms. Hence the time-interval differences used in that previous study was on the same order of the timing changes in AF induced by the Accel and Decel perturbations used in the current one. It was found that the group of 30 PWS made significantly more errors (by about 4 percentage points) than matched controls. These results indicated that stutterer do have deficits in sub-second timing processing, perhaps in both perception and production, due to the shared neural mechanisms. This may be related to the finding of the current study that PWS were less able to react in a compensatory manner to timing changes on the order of tens of milliseconds in auditory feedback during speech articulation.

The detailed neural underpinning of the millisecond motor timing deficit in PWS is elusive at this moment. It is widely believed that the cerebellum is a critical neural structure in millisecond timing (Buhusi and Meck 2005; Koch et al. 2007). In this regard, it is noteworthy that a previous study found GM volume reduction in the cerebellum in PWS (based on VBM, Song et al. 2007) and another previous study showed abnormal functional connectivity between the right

cerebellum and cerebral cortical areas, including the bilateral precentral gyri and the right angular gyrus (Lu et al . 2009). Also, it is worth pointing out that most previous behavioral studies investigated millisecond motor timing by using relatively simple tasks such as finger tapping. The nature of these simple actions may belie the true neural substrate of millisecond motor timing in speech production, which involves much more complex movements and richer sensory contextual information. It is likely that for millisecond motor timing in such complex sensorimotor actions, several other brain regions, such as the basal ganglia, work with the cerebellum (Wildgruber et al. 2001; Ackermann et al. 2008).

Finally, some discussion is warranted with regard the relations between the different patterns of behavioral abnormalities found under the static-vowel perturbation (Experiment A) and the time-varying perturbations (Experiment B and C). In Experiment A, we observed that the PWS showed significantly smaller compensation magnitude at 300 ms (a relatively long period of time) following the onset of the perturbation. However, in Experiments B and C, the data showed that whereas the PWS made significantly smaller-than-normal corrective adjustments in early part of the compensation, their later compensatory responses did not differ from normal. These patterns of between-group differences may seem different at first glance, because the first difference appears to be a sustained difference, whereas the latter appears to be a transient one. However, this discrepancy here may be attributable to the different time-coureses of the perturbations involved in the two types of experiments. The static-vowel perturbation involved a sudden-onset, longer-lasting, and nearly constant-magnitude shifting of AF, which may constantly tax the capacity of the IMs to generate counteracting motor commands. By contrast, the perturbations imposed on the time-varying utterance were transient and contained smooth on- and off-ramps, and therefore may have imposed a relatively short-lasting and weaker disruption of the normal time-course of sensorimotor processes. Due to this relatively more fleeting nature of the time-varying perturbation, the IMs of the PWS may have had sufficient time to ramp up the perturbation in later part of the compensation, which may have rendered the between-group differences in the late responses non-significant. Although this explanation for the discrepant

findings from the static and time-varying perturbation experiments is somewhat speculative and certainly awaits future tests and confirmation, it is consistent with the unpublished findings from Ludo Max and his colleagues that in an auditory-motor adaptation experiment based on sustained AF perturbation, the PWS made slower-than-normal adaptive corrections to their productions in early phases of the adaptation, but given enough trials, the PWS' adaptive responses became closer to (i.e., caught up with) the response of the normal controls (L. Max, personal communication). Hence, there seems to be converging evidence from these unpublished adaptation results and our findings that the IMs in the speech motor system of a PWS are not completely defective, but the primary deficit is a sluggishness in performance compared the IMs of a normally fluent person in performing functions such as generating feedback-based motor corrections, self-updating based on feedback errors. This slower-than-normal performance may also be applicable to the function of generating motor commands based on auditory targets or phonemic sequence information online (i.e., feedforward control). This could be an explanation for the rate effect in stuttering: the level of dysfluency of PWS tend to decrease with slower speaking rates (e.g., Adams et al. 1973), the slowing down of the speaking rate provides more time between syllables, hence allowing the IMs to generate appropriate motor programs for ensuring the accurate and fluent production of the syllable sequences.

Chapter 5. Summary of findings and future directions

5.1. Summary of main findings and results

The following is a brief summary of the main findings and results of this dissertation.

1) In Chapter 2, we described a novel experimental method based on time-varying AF perturbation to examine the role of AF in the online control of multisyllabic articulation. Under a type of perturbations that manipulated the magnitude of formants in the time-varying trajectory, a group of 36 normal subjects made small but significant compensatory alterations to their produced formant trajectories, indicating that AF is utilized by the speech motor system to control the spatial (position) parameters of multisyllabic articulation. These compensatory formant changes could be observed not only in the syllable under perturbation, but also in the syllables that were produced after the cessation of the perturbation, indicating an online control scheme in which AF state information from preceding articulatory units is utilized to guide or fine-tune the production of later ones.

   In addition, under a second type of AF perturbation that manipulated the pace of the formant trajectory evolution, 29 normal subjects showed an asymmetric pattern of timing adjustment. They showed relatively small amount of timing changes under the accelerating perturbation, but showed much greater timing adjustments in the direction of slowing down the production under the decelerating perturbation. These findings provide the first unequivocal support for the involvement of AF in the online control of articulatory timing. These data provide a glance into the complexity of the sensorimotor processes underlying the production of connected speech and provide constraints for computational models of such processes.

2) In Chapter 3 of this dissertation, we described sqDIVA, a computational model we developed of the sensorimotor processes underlying the control of the articulation of syllable sequences and between-syllable transitions during multisyllabic speech

utterances. sqDIVA is a model that is kept intentionally simple by making a number of assumptions and omissions, including the omission of the details of vocal-tract geometry and articulatory-acoustic transformation, a simple timing planning regime, as well as the focus on a single output variable (F2). However, in this model, we carved out the mathematical details of the way in which the speech motor system utilizes AF to perform online fine-tuning of the spatial and temporal parameters of multisyllabic articulation and sequencing. The experimental data from Chapter 2 were used in fitting the model and validating it against a Baseline model as well as another model with the same number of degrees of freedom. Modeling results indicated that the simple concept of updating syllable onset and offset timing based on the auditory state could provide a unifying explanation for the spatiotemporal compensation patterns observed in Chapter 2. These results provide strong support for online AF-based timing adjustment for being an independent and important component of online speech motor control.

3) In Chapter 4, we described a systematic investigation of the difference between stutterers (PWS) and nonstutterers (PFS) in the online AF-based control of articulation. Three experiments, which covered the schemes of both quasi-static and time-varying articulation and both spatial and temporal parameter control, were administered to groups of PWS and PFS. The results from these three experiments showed a converging pattern of significantly weaker-than-normal compensatory responses by PWS. Based on the theoretical assumption that this type of online auditory-motor adaptation relies on the internal inverse and forward modeling, these between-group differences provide support for the hypothesis that stuttering involves an abnormally learned or activated set of IMs in the speech motor system. In particular, the later-than-normal onset of the significant compensatory responses in the PWS under the time-varying perturbations seem to indicate that the speech motor IMs are not completely dysfunctional in stuttering, but are instead significantly slower than normal in computing or implementing proper corrective motor commands based on AF error information. The implication of this new behavioral

finding for the neural mechanisms of stuttering and its relation to what we already know about the various behavioral characteristics of this disorder was described at the end of Chapter 4.

5.2. Limitations and future directions

First, due to methodological considerations, the stimulus utterance we used in Chapter 2 was limited in that it consisted of only semivowels and vowels. The applicability of the conclusions based on the findings of Chapter 2 needs to be confirmed in future studies that use stimulus utterances with more generic phonemic compositions. To this end, changes and improvements will need to be to the Audapter platform.

Many important questions follow the findings of Chapter 2 regarding the auditory-motor interaction during multisyllabic speech production. One of the most important questions is the underlying neural mechanisms for the AF-based timing adjustment. This question can be investigated by using fMRI that employs sparse-sampling event-related designs similar to that of Tourville et al. (2008) or with transcranial magnetic stimulation. The former method will be instrumental in mapping out the anatomical locations of the neural activities underlying the detection of and compensation to the online AF errors, which we hypothesize to be similar to the brain regions reported by Tourville et al. (2008) and Golfinopoulos et al. (2011), i.e., the involvement of the right vPMC and cerebellum. However, we further hypothesize that due to the timing component of this feedback-based control, the supplementary motor area and the associated BG-thalamic loop will also show increased activation under the Decel-type timing perturbation. Unlike fMRI, TMS (e.g., Maeda and Pascual-Leone 2003) will enable us to establish the causal roles of given brain regions in this feedback-based control through temporary disruption of neural activities. We hypothesize that single-pulse stimulation of the SMA within a tight time window the time around the time of the temporal AF perturbation should reduce the magnitude of the timing adjustment.

Another interesting question regarding the generality of the feedback-based motor timing control is the applicability of the control scheme embodied by the sqDIVA model to modes of vocalization other than naturally timed speech (as used in Chapters 2 and 4). Is it possible that under rhythmic timing of speech (such as chanting and singing), the role of AF in the online motor timing control is diminished compared to natural speech? This question has important implications for the fluency-enhancing effect of rhythmic pacing in stuttering, because if AF-based timing control mode is bypassed during rhythmic speech, it may provide support an important role of deficits in AF-based syllable triggering in the etiology of stuttering (see Fig. 4.18), which may lead to important questions regarding the neural substrates of this deficit.

The sqDIVA model we developed in Chapter 3 is largely a mid-level conceptual model that is meant to be an important piece of the link between DIVA, a model for low-level, single-syllable sensorimotor control, and GODIVA, a model for high-level, cognitive-linguistic planning and sequencing of the phonemic and syllabic content in connected speech. We plan to integrate DIVA and GODIVA, with sqDIVA as an important bridge, in the near future. There are some important theoretical and modeling questions to be addressed in this integration process. For example, how are motor trajectories for the between syllable transitions generated? Are they preplanned or are they generated on the fly? How does inverse IMs participate in this process of generating transitional trajectories? How should one model the effect of speaking rate changes in the integrated DIVA-GODIVA framework? How should one account for anticipatory coarticulation across syllabic boundaries? Is it possible to devise a neurophysiologically more plausible mechanism for generating timing scores for multisyllabic speech utterances (e.g., see Fig. 3.1.A) And moreover, can we make specific and well-motivated modifications to the parameters of this integrative model to generate stuttering-like behavior? These are all interesting and challenging questions for future modeling efforts.

With respect to the finding of weakened auditory-motor compensation during speech articulation in PWS, questions about the underlying neural substrates arise. We are currently pursuing systematic investigation of white-matter integrity in stuttering with structural MRI and

DTI and how it may be correlated with the under-compensation described in Chapter 4. Two possibilities regarding abnormal WM deficits arise according to previous neurophysiological findings. First, it is possible that the deficits in the right arcuate fasciculus may disrupt the normal communication between the auditory areas and the right vPMC, regarded as the location of the feedback control map (Golfinopoulos et al. 2010). In addition, it is possible that WM connections between cerebellum and other parts of the brain may lead to deficits in the learning, updating and activation of sensorimotor IMs.

Alternative explanations exist for the auditory-motor under-compensation we found in Chapter 4. For example, since adult PWS were used in our study, it is conceivable that this under-utilization of AF for online speech motor control is a compensatory strategy developed through extensive experience with a defective speech motor network in order to reduce the reliance on an unstable feedback control system, rather than reflecting intrinsic deficits of the sensorimotor IMs. In other words, the causal relation between the core characteristics of stuttering and the abnormal auditory-motor interaction found in the current study remains elusive. This question can be partially addressed by performing online AF perturbation experiment similar to the one used in Chapter 4 to children who stutter. Because stuttering is by its nature a developmental disorder that has roots in early speech motor development (Bloodstein and Ratner 2008), elucidating the sensorimotor deficits at an early stage of the development and its relations to neuroanatomical properties of the brain will be invaluable in bringing us closer to understanding the etiology and pathophysiology of this disorder.

Bibliography

Abbs JH, Gracco VL (1983) Sensorimotor Actions in the Control of Multi-Movement Speech Gestures. Trends in Neurosciences 6:391-395.

Abbs JH, Gracco VL (1984) Control of Complex Motor Gestures - Orofacial Muscle Responses to Load Perturbations of Lip during Speech. J Neurophysiol 51:705-723.

Ackermann H (2008) Cerebellar contributions to speech production and speech perception: Psycholinguistic and neurobiological perspectives. Trends Neurosci 31:265-272.

Adams MR, Lewis JI, Besozzi TE (1973) The effect of reduced reading rate on stuttering frequency. J Speech Hear Res 16:671-675.

Adams MR, Hayden P (1976) The ability of stutterers and nonstutterers to initiate and terminate phonation during production of an isolated vowel. J Speech Hear Res 19:290-296.

Ajemian R, Hogan N (2010) Experimenting with theoretical motor neuroscience. J Mot Behav 42:333-342.

Ambrose NG, Cox NJ, Yairi E (1997) The genetic basis of persistence and recovery in stuttering. J Speech Lang Hear Res 40:567-580.

Amunts K, Schleicher A, Burgel U, Mohlberg H, Uylings HB, Zilles K (1999) Broca's region revisited: cytoarchitecture and intersubject variability. J Comp Neurol 412:319-341.

Armson J, Stuart A (1998) Effect of extended exposure to frequency-altered feedback on stuttering during reading and monologue. J Speech Lang Hear R 41:479-490.

Ashburner J, Friston KJ (2000) Voxel-based morphometry--the methods. Neuroimage 11:805-821.

Atkinson CJ (1953). Adaptation to delayed sidetone. J Speech Hear Disord 18:386-391.

Barasch CT, Guitar B, McCauley RJ, Absher RG (2000) Disfluency and time perception. J Speech Lang Hear Res 43:1429-1439.

Barlow SM, Andreatta RD (1999) Handbook of clinical speech physiology. San Diego: Singular Pub. Group.

Bauer A, Jancke L, Kalveram KT (1997) Mechanical perturbation of the jaw during speech in stutterers and nonstutterers. In: Speech Production: Motor Control, Brain Research and Fluency Disorders(Hulstijn, W. et al., eds), pp 191-196 Amsterdam: Elsevier Science.

Bauer JJ, Larson CR (2003) Audio-vocal responses to repetitive pitch-shift stimulation during a sustained vocalization: Improvements in methodology for the pitch-shifting technique. Journal of the Acoustical Society of America 114:1048-1054.

Beal DS, Cheyne DO, Gracco VL, Quraan MA, Taylor MJ, De Nil LF (2010) Auditory evoked fields to vocalization during passive listening and active generation in adults who stutter. Neuroimage 52:1645-1653.

Beal DS, Quraan MA, Cheyne DO, Taylor MJ, Gracco VL, De Nil LF (2011) Speech-induced suppression of evoked auditory fields in children who stutter. Neuroimage 54:2994-3003.

Behrens TE, Berg HJ, Jbabdi S, Rushworth MF, Woolrich MW (2007) Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? Neuroimage 34:144-155.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 57:289-300.

Black JW (1951) The effect of delayed sidetone upon vocal rate and intensity. J speech Hear Disord 16:56-60.

Bloodstein O, Ratner NB (2008) A handbook on stuttering. Clifton Park, NY: Thomson/Delmar Learning.

Bohland JW, Bullock D, Guenther FH (2009) Neural Representations and Mechanisms for the Performance of Simple Speech Sequences. J Cogn Neurosci. 22(7):1504-1529.

Bohland JW, Guenther FH (2006) An fMRI investigation of syllable sequence production. Neuroimage 32:821-841.

Borden GJ (1979) Interpretation of Research on Feedback Interruption in Speech. Brain and Language 7:307-319.

Boone D (1966). Modification of the voices of deaf children. Volta Rev 68:686-692.

Boucek M (2007) The nature of planned acoustic trajectories. Unpublished M.S. thesis: University Karlsruhe

Braun AR, Varga M, Stager S, Schulz G, Selbie S, Maisog JM, Carson RE, Ludlow CL (1997) Altered patterns of cerebral activity during speech and language production in developmental stuttering. An H2(15)O positron emission tomography study. Brain 120 (Pt 5):761-784.

Browman CP, Goldstein L (1992) Articulatory phonology: an overview. Phonetica 49:155-180.

Brown SF (1938) The theoretical importance of certain factors influencing the incidence of stuttering. J Speech Disord 3:223-230.

Buhusi CV, Meck WH (2005) What makes us tick? Functional and neural mechanisms of interval timing. Nat Rev Neurosci 6:755-765.

Bullock D, Rhodes BJ (2003) Competitive queuing for planning and serial performance. In: The Handbook of Brain Theory and Neural Networks (2nd Ed) (Arbibi, M. A., ed), pp 241-244 Cambridge, MA: MIT Press.

Burnett TA, Freedland MB, Larson CR, Hain TC (1998) Voice F0 responses to manipulations in pitch feedback. Journal of the Acoustical Society of America 103:3153-3161.

Burnett TA, Larson CR (2002) Early pitch-shift response is active in both steady and dynamic voice pitch control. Journal of the Acoustical Society of America 112:1058-1063.

Cai S, Beal DS, Tiede MK, Perkell JS, Guenther FH, and Ghosh SS. (2011). Relating the kinematic variability of speech to MRI-based structural integrity of brain white matter in people who stutter and people with fluent speech. To be presented at Society for Neuroscience Annual Meeting 2011, Washington, DC, Nov. 12 – 16.

Cai S, Ghosh SS, Guenther FH, Perkell JS (2010)a Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iau/ and its pattern of generalization. J Acoust Soc Am 128:2033-2048.

Cai S, Ghosh SS, Guenther FH, Perkell JS (2010b). Coordination of the first and second formants of the Mandarin triphthong /iau/ revealed by adaptation to auditory perturbations (Abstract). J Acoust Soc Am 127:2018.

Calvert D (1961). Deaf voice quality: a preliminary investigation. Volta Rev 64:402-403.

Caruso AJ, Gracco VL, Abbs JH (1987) A speech motor control perspective on stuttering: Preliminary observations. In: Speech motor dynamics in stuttering(Peters, H. F. M. and Hulstijn, W., eds), pp 245-258 Wien, Austria: Springer-Verlag.

Chang SE (2011) Using Brain Imaging to Unravel the Mysteries of Stuttering. In: The Dana Foundation. Data accessed: 11/05/2011. URL: http://dana.org/news/cerebrum/detail.aspx?id=33796

Chang SE, Erickson KI, Ambrose NG, Hasegawa-Johnson MA, Ludlow CL (2008) Brain anatomy differences in childhood stuttering. Neuroimage 39:1333-1344

Chang SE, Horwitz B, Ostuni J, Reynolds R, Ludlow CL (2011) Evidence of left inferior frontal-premotor structural and functional connectivity deficits in adults who stutter. Cereb Cortex. 21(11): 2507-2518.

Chen SH, Liu H, Xu Y, Larson CR (2007) Voice F0 responses to pitch-shifted voice feedback during English speech. J Acoust Soc Am 121:1157-1163.

Chen Z, Liu P, Jones JA, Huang D, Liu H (2010) Sex-related differences in vocal responses to pitch feedback perturbations during sustained vocalization. J Acoust Soc Am 128:EL355-EL360.

Cherry C, Sayers BM (1956) Experiments upon the total inhibition of stammering by external control. J Psychosom Res 1:223-246.

Christoffels IK, Formisano E, Schiller NO (2007) Neural correlates of verbal feedback processing: an fMRI study employing overt speech. Hum Brain Mapp 28:868-879.

Civier O, (2010). Computational modeling of the neural substrates of stuttering and induced fluency (Doctoral dissertation, Boston University, 2010). Dissertation Abstracts International, 70, 10.

Civier, Bullock, Max & Guenther (submitted for publication) Computational modeling of stuttering caused by impairments in a basal ganglia thalamo-cortical circuit involved in syllable selection and initiation.

Civier O, Tasko SM, Guenther FH (2010) Overreliance on auditory feedback may lead to sound/syllable repetitions: simulations of stuttering and fluency-inducing conditions with a neural model of speech production. J Fluency Disord In press.

Conture EG, Brayton ER (1975) Influence of Noise on Stutterers Different Disfluency Types. J Speech Hear Res 18:381-384.

Craig A, Tran Y (2005) The epidemiology of stuttering: the need for reliable estimates of prevlance and anxiety levels over the lifespan. Advances in Speech Language Pathology 7:41-46.

Cross DE, Luper HL (1983) Relation between finger reaction time and voice reaction time in stuttering and nonstuttering children and adults. J Speech Hear Res 26:356-361.

Culton GL (1986) Speech disorders among college freshmna: A 13-year survey. Journal of Speech and Hearing Disorders 51:3-7.

Curio G, Neuloh G, Numminen J, Jousmaki V, Hari R (2000) Speaking modifies voice-evoked activity in the human auditory cortex. Human Brain Mapping 9:183-191.

Crystal TH, House AS (1998) Segmental durations in connected speech signals: Current results. J Acoust Soc Am 83:1553-1573.

Cowie R, Douglas-Cowie E, Kerr AG (1982) A study of speech deterioration in post-lingually deafened adults. J Laryngol Otol 96:101-112.

Cykowski MD, Fox PT, Ingham RJ, Ingham JC, Robin DA (2010) A study of the reproducibility and etiology of diffusion anisotropy differences in developmental stuttering: a potential role for impaired myelination. Neuroimage 52:1495-1504.

Davidson GD (1959). Sidetone delay and reading rate, articulation, and pitch. J Speech Hear Res 2:266-270.

Daniloff R, Hammarberg R (1973) On defining coarticulation. J Phonetics 1:239-248.

De Nil LF, Kroll RM, Kapur S, Houle S (2000) A positron emission tomography study of silent and oral single word reading in stuttering and nonstuttering adults. J Speech Lang Hear R 43:1038-1053.

Donath TM, Natke U, Kalveram KT (2002) Effects of frequency-shifted auditory feedback on voice F0 contours in syllables. Journal of the Acoustical Society of America 111:357-366.

Eliades SJ, Wang X (2008) Neural substrates of vocalization feedback monitoring in primate auditory cortex. Nature 453:1102-1106.

Elman JL (1981) Effects of frequency - shifted feedback on the pitch of vocal productions. J Acoust Soc Am 70:45-50.

Ezrati-Vinacour R, Levin I (2001) Time estimation by adults who stutter. J Speech Lang Hear Res 44:144-155.

Flash T, Hogan N (1985) The Coordination of Arm Movements - an Experimentally Confirmed Mathematical-Model. J Neurosci 5:1688-1703.

Fairbanks G (1955). Selective vocal effects of delayed auditory feedback. J Speech Hear Disord 20:333-345.

Farnetani E, Recasens D (1999) Coartculation models in recent speech production theories. In: Coarticulation: Theory, Data and Techniques(Hardcastle, W. J. and Hewlett, N., eds), pp 31-65 Cambridge, UK: Cambridge University Press.

Fairbanks G, Guttman N (1958) Effects of Delayed Auditory-Feedback Upon Articulation. J Speech Hear Res 1:12-22.

Feldman AG (1966) Function tuning of the nervous system with control of movement of maintenance of a steady posture: II. Control paramters of the muscle. Biophysics 11:565-578.

Feldman AG (1986) Once more on the equilibrium-point hypothesis (lambda model) for motor control. J Mot Behav 18:17-54.

Feng Y, Gracco VL, Max L (2011) Integration of auditory and somatosensory error signals in the neural control of speech movements. J Neurophysiol 106:667-679.

Ferrand CT, Gilbert HR, Blood GW (1991) Selected aspects of central processing and vocal motor function in stutterers and nonstutterers : P300, laryngeal shift, and vibratory onset J Fluency Disord 16:101-115.

Folkins JW, Abbs JH (1975) Lip and Jaw Motor Control during Speech - Responses to Resistive Loading of Jaw. J Speech Hear Res 18:207-220.

Fowler CA (1980) Coarticulation and Theories of Extrinsic Timing. Journal of Phonetics 8:113-133.

Foundas AL, Bollich AM, Feldman J, Corey DM, Hurley M, Lemen LC, Heilman KM (2004) Aberrant auditory processing and atypical planum temporale in developmental stuttering. Neurology 63:1640-1646.

Fox PT, Ingham RJ, Ingham JC, Hirsch TB, Downs JH, Martin C, Jerabek P, Glass T, Lancaster JL (1996) A PET study of the neural systems of stuttering. Nature 382:158-161.

Fox PT, Ingham RJ, Ingham JC, Zamarripa F, Xiong JH, Lancaster JL (2000) Brain correlates of stuttering and syllable production. A PET performance-correlation analysis. Brain 123 (Pt 10):1985-2004.

Gold T (1980). Speech production in hearing-impaired children. J Commun Disord 13:397-418.

Golfinopoulos E, Tourville JA, Guenther FH (2009) The integration of large-scale neural network modeling and functional brain imaging in speech motor control. Neuroimage 52:862-874.

Golfinopoulos E, Tourville JA, Bohland JW, Ghosh SS, Nieto-Castanon A, Guenther FH (2011). FMRI investigation of unexpected somatosensory feedback perturbation during speech. Neuroimage 55:1324-1338.

Gould J, Lane H, Vick J, Perkell JS, Matthies ML, Zandipour M (2001). Changes in speech intelligibility of postlingually deaf adults after cochlear implantation. Ear Hear 22:453-460.

Gracco VL, Abbs JH (1985) Dynamic control of the perioral system during speech: kinematic analyses of autogenic and nonautogenic sensorimotor processes. J Neurophysiol 54:418-432.

Gracco VL, Abbs JH (1989) Sensorimotor characteristics of speech motor sequences. Exp Brain Res 75:586-598.

Guenther FH (1995) Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychol Rev 102:594-621.

Guenther FH, Espy-Wilson CY, Boyce SE, Matthies ML, Zandipour M, Perkell JS (1999) Articulatory tradeoffs reduce acoustic variability during American English /r/ production. J Acoust Soc Am 105:2854-2865.

Guenther FH, Hampson M, Johnson D (1998) A theoretical investigation of reference frames for the planning of speech movements. Psychol Rev 105:611-633.

Guenther FH (2006).Cortical interactions underlying the production of speech sounds. J Commun Disord 39:350-365.

Hall JW, Jerger J (1978) Central auditory function in stutterers. J Speech Hear Res 21:324-337.

Hanna R, Morris S (1977) Stuttering, speech rate, and the metronome effect. Percept Motor Skills 44:452-454.

Harrington J (1988). Stuttering, delayed auditory feedback, and linguistic rhythm. J Speech Hear Res 31:36-47.

Heinks-Maldonado TH, Nagarajan SS, Houde JF (2006) Magnetoencephalographic evidence for a precise forward model in speech production. Neuroreport 17:1375-1379. Helm NA, Butler RB, Benson DF (1978) Acquired stuttering. Neurology 28:1159-1165.

Herndon GY (1966) A study of the time discrimination abilities of stutterers and nonstutterers., vol. Doctoral dissertation: University of Oklahoma.

Hickok G, Houde JF, Rong F (2011) Sensorimotor integration in speech processing: Computational basis and neural organization. Neuron 69:407-422.

Honda M, Akinori F, Kaburagi T (2002) Compensatory responses of articulators to unexpected perturabtion of the palate shape. J Phonetics 30:281-302.

Houde JF, Jordan MI (1998) Sensorimotor adaptation in speech production. Science 279:1213-1216.

Houde JF, Jordan MI (2002) Sensorimotor adaptation of speech I: Compensation and adaptation. J Speech Lang Hear R 45:295-310. Hudgins C, Numbers F (1942) An investigation of the intelligibility of the speech of the deaf. Am Ann Deaf 82:338-363.

Houde JF, Nagarajan SS, Sekihara K, Merzenich MM (2002) Modulation of the auditory cortex during speech: an MEG study. J Cogn Neurosci 14:1125-1138.

Howell P (2004) Assessment of Some Contemporary Theories of Stuttering That Apply to Spontaneous Speech. Contemp Issues Commun Sci Disord 31:122-139.

Howie PM (1981) Concordance for Stuttering in Monozygotic and Dizygotic Twin Pairs. J Speech Hear Res 24:317-321.

Hudgins C, Numbers F (1942) An investigation of the intelligibility of the speech of the deaf. Am Ann Deaf 82:338-363.

Huettel SA, Song AW, McCarthy G (2008) Functional Magnetic Resonance Imaging: Sinauer Associates, Inc.

Hutchinson JM, Norris GM (1977) The differential effect of three auditory stimuli on the frequency of stuttering behaviors. J Fluency Disord 2:283-293.

Indefrey P, Levelt WJ (2004) The spatial and temporal signatures of word production components. Cognition 92:101-144.

Ingham RJ, Moglia RA, Frank P, Ingham JC, Cordes AK (1997) Experimental investigation of the effects of frequency-altered auditory feedback on the speech of adults who stutter. J Speech Lang Hear Res 40:361-372.

Ivry RB, Hazeltine RE (1995) Perception and Production of Temporal Intervals across a Range of Durations - Evidence for a Common Timing Mechanism. J Exp Psychol Human 21:3-18.

Jones JA, Munhall KG (2002) The role of auditory feedback during phonation: studies of Mandarin tone production. Journal of Phonetics 30:303-320.

Jones RD, White AJ, Lawson KH, Anderson TJ (2002) Visuoperceptual and visuomotor deficits in developmental stutterers: an exploratory study. Hum Mov Sci 21:603-619.

Kalinowski J, Armson J, Roland-Mieszkowski M, Stuart A, Gracco VL (1993) Effects of alterations in auditory feedback and speech rate on stuttering frequency. Lang Speech 36 (Pt 1):1-16.

Kalveram KT (1991) How pathological audio-phonatoric coupling induces stuttering: A model of speech flow control. In: Speech Motor Control and Stuttering (Peters, H. F. M. et al., eds), pp 163-170 Amsterdam: Elsevier Science.

Kalveram KT, Jancke L (1989) Vowel duration and voice onset time for stressed and nonstressed syllables in stutterers under delayed auditory feedback condition. Folia Primatologica 41:30-42.

Kang C, Riazuddin S, Mundorff J, Krasnewich D, Friedman P, Mullikin JC, Drayna D (2010) Mutations in the lysosomal enzyme-targeting pathway and persistent stuttering. N Engl J Med 362:677-685.

Kawato M (1999) Internal models for motor control and trajectory planning. Current Opinion in Neurobiology 9:718-727.

Keating PA (1988) The window model of coartciulation: articulatory evidence. UCLA Working Papers in Phonetics 69:3-29.

Keele SW (1968) Movement control in skilled motor performance. Psychol Bull 70:387-403.

Keele SW, Pokorny RA, Corcos DM, Ivry R (1985) Do Perception and Motor Production Share Common Timing Mechanisms - a Correlational Analysis. Acta Psychol 60:173-191.

Kell CA, Neumann K, von Kriegstein K, Posenenske C, von Gudenberg AW, Euler H, Giraud AL (2009) How the brain repairs stuttering. Brain 132:2747-2760.

Kent RD (1984) Stuttering as a temporal programming disorder. In: Nature and Treatment of Stuttering(Curlee, R. F. and Perkins, W. H., eds), pp 283-301 Boston, MA: Allyn and Bacon.

Kidd KK, R. KJ, Records MA (1978) The possible causes of sex ratio in stuttering and its implications. J Fluency Disord 3:13-23.

Klatt DH (1980) Software for a cascade/parallel formant synthesizer. J Acoust Soc Am 67:971-995.

Kleinow J, Smith A (2000) Influences of length and syntactic complexity on the speech motor stability of the fluent speech of adults who stutter. J Speech Lang Hear Res 43:548-559.

Klich RJ, May GM (1982) Spectrographic study of vowels in stutterers' fluent speech. . J Speech Hear Res 25:364-370.

Koch G, Oliveri M, Torriero S, Salerno S, Lo Gerfo E, Caltagirone C (2007) Repetitive TMS of cerebellum interferes with millisecond time processing. Exp Brain Res 179:291-299.

Korowbow N (1955). Reaction to stress: a reflection of personality trait organization. J Abnorm Soc Psychol 51:464-468.

Lane H, Denny M, Guenther FH, Hanson HM, Marrone N, Matthies ML, Perkell JS, Stockmann E, Tiede M, Vick J, Zandipour M (2007). On the structure of phoneme categories in listeners with cochlear implants. J Speech Lang Hear Res 50:2-14.

Lane H, Wozniak J, Perkell J (1994) Changes in voice-onset time in speakers with cochlear implants. Journal of the Acoustical Society of America 96:56-64.

Lane H, Tranel B (1971) Lombard Sign and Role of Hearing in Speech. Journal of Speech and Hearing Research 14:677-&.

Lane H, Tranel B, Sisson C (1970). Regulation of voice communication by sensory dynamics. J Acoust Soc Am 47:618-624.

Larson CR, Altman KW, Liu HJ, Hain TC (2008) Interactions between auditory and somatosensory feedback for voice F-0 control. Exp Brain Res 187:613-621.

Larson CR, Burnett TA, Kiran S, Hain TC (2000) Effects of pitch-shift velocity on voice F-0 responses. Journal of the Acoustical Society of America 107:559-564.

Lee BS (1950) Effects of delayed auditory feedback. J Acoust Soc Am 22:824-826.

Lee BS (1951) Artificial stutter. J Speech Disord 16:53-55.

Levelt WJM (1989) Speaking: from intention to articulation. Cambridge, Mass.: MIT Press.

Levelt WJM, Wheeldon L (1994) Do Speakers Have Access to a Mental Syllabary. Cognition 50:239-269.

Levitt H (1970) Transformed up-down methods in psychophysics. J Acoust Soc Am 49:467-477.

Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A (2001) Neurophysiological investigation of the basis of the fMRI signal. Nature 412:150-157.

Loucks TM, De Nil LF (2006) Oral kinesthetic deficit in adults who stutter: a target-accuracy study. J Mot Behav 38:238-246.

Loucks TM, De Nil LF, Sasisekaran J (2007) Jaw-phonatory coordination in chronic developmental stuttering. J Commun Disord 40:257-272.

Lu C, Ning N, Peng D, Ding G, Li K, Yang Y, Lin C (2009) The role of large-scale neural interactions for developmental stuttering. Neuroscience 161:1008-1026.

Ludlow CL, Loucks T (2003) Stuttering: a dynamic motor control disorder. J Fluency Disord 28:273-295; quiz 295.

Munhall KG, Lofqvist A, Kelso JAS (1994) Lip-Larynx Coordination in Speech - Effects of Mechanical Perturbations to the Lower Lip. J Acoust Soc Am 95:3605-3616.

MacDonald EN, Goldberg R, Munhall KG (2010) Compensations in response to real-time formant perturbations of different magnitudes. J Acoust Soc Am 127:1059-1068.

MacNeilage PF (1998) The frame/content theory of evolution of speech production. Behav Brain Sci 21:499-511; discussion 511-446.

Maeda F, Pascual-Leone A (2003) Transcranial magnetic stimulation: studying motor neurophysiology of psychiatric disorders. Psychopharmacology (Berl) 168:359-376.

Maeda S (1982) A digitial simulation method of the vocal-tract system. Speech Commun 1:199-229.

Mansson H (2000) Childhood stuttering: incidence and development. J Fluency Disord 25:47-57.

Maraist JA, Hutton C (1957) Effects of auditory masking upon the speech of stutterers. J Speech Hear Disord 22:385-389.

Markides A (1970). The speech of deaf and partially-hearing children with special reference to factors affect intelligibility. Brit J Dis Commun 5:126-140.

Martin RR, Johnson LJ, Siegel GM, Haroldson SK (1985) Auditory stimulation, rhythm, and stuttering. J Speech Hear Res 28:487-495.

Martin RR, Siegel GM, Johnson LJ, Haroldson SK (1984) Sidetone amplification, noise, and stuttering. J Speech Hear Res 27:518-527.

Matthies ML, Svirsky M, Perkell J (1996) Acoustic and articulatory measures of sibilant production with and without auditory feedback from a cochlear implant. J Speech Lang Hear Res 39:936-946.

Matthies ML, Guenther FH, Denny M, Perkell JS, Burton E, Vick J, Lane H, Tiede M, Zandipour M (2008) Perception and production of /r/ allophones improve with hearing from a cochlear implant. J Acoust Soc Am 124:3191-3202.

Max L, Baldwin CJ (2010) The role of motor learning in stuttering adaptation: repeated versus novel utterances in a practice-retention paradigm. J Fluency Disord 35:33-43.

Max L, Guenther FH, Gracco VL, Ghosh S (2004). Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: A theoretical model of stuttering. Contemporary Issues in Communication Science and Disorders 31:105-122.

Ménard L, Polak M, Denny M, Burton E, Lane H, Matthies ML, Marrone N, Perkell JS, Tiede M, Vick J (2007). Interactions of speaking condition and auditory feedback on vowel production in postlingually deaf adults with cochlear implants. Journal of the Acoustical Society of America 121:3790-3801.

Miall RC, Wolpert DM (1996) Forward models for physiological motor control. Neural Networks 9:1265-1279.

Mochida T, Gomi H, Kashino M (2010) Rapid change in articulatory lip movement induced by preceding auditory feedback during production of bilabial plosives. PLoS ONE 5:e13866.

Munhall KG, MacDonald EN, Byrne SK, Johnsrude I (2009) Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. J Acoust Soc Am 125:384-390.

Mysak ED (1959) A servo model for speech therapy. J Speech Hear Disord 24:144-149.

Mysak ED (1960) Servo Theory and Stuttering. Journal of Speech and Hearing Disorders 25:188-195.

Nam H, Saltzman E (2003) A competitive, coupled oscillator model of syllable structure. In: Proceedings of the XVth International Congress of Phonetic Sciences Barcelona, Spain.

Namasivayam AK, van Lieshout P, McIlroy WE, De Nil L (2009).Sensory feedback dependence hypothesis in persons who stutter. Hum Mov Sci 28:688-707.

Nasir SM, Ostry DJ (2008) Speech motor learning in profoundly deaf adults. Nat Neurosci 11:1217-1222.

Natke U, Donath TM, Kalveram KT (2003) Control of voice fundamental frequency in speaking versus singing. Journal of the Acoustical Society of America 113:1587-1593.

Natke U, Kalveram KT (2001) Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables. J Speech Lang Hear R 44:577-584.

Neilson MD, Neilson PD (1979) Systems analysis of tracking performance in stutterers and normals. (Abstract). In: American Speech and Hearing Association, vol. 21, p 770.

Neilson MD, Neilson PD (1987) Speech Motor Control And Stuttering - A Computational Model Of Adaptive Sensory-Motor Processing. Speech Commun 6:325-333.

Niziolek CA (2010) The role of linguistic contrasts in the auditory feedback control of Speech. Ph.D. dissertation: Massachusetts Institute of Technology, Cambridge, MA, USA.

Nober E (1967) Articulation of the deaf. Excep Child 33:611-621.

Norman DA (1980) Copycat science or does the mind really work by table look-up. In: Perception and Production of Fluent Speech(Cole, R. A., ed) Hillsdale, NJ: Lawrence Erlbaum.

Nudelman HB, Herbrich KE, Hess KR, Hoyt BD, Rosenfield DB (1992) A Model of the Phonatory Response-Time of Stutterers and Fluent Speakers to Frequency-Modulated Tones. J Acoust Soc Am 92:1882-1888.

Osberger MJ, Maso M, Sam LK (1993) Speech intelligibility of children with cochlear implants, tactile aids, or hearing aids. Journal of Speech and Hearing Research 36:186-203.

Parkhurst B, Levitt H (1978). The effect of selected prosodic errors on the intelligibility of deaf speech. J Commun Disord 11:249-256.

Öhman S (1966) Coarticulation in VCV utterances: spectrographic measurements. J Acoust Soc Am 39:151-168.

Öhman S (1967) Numerical model of coarticulation. J Acoust Soc Am 41:310-320.

Paus T, Perry DW, Zatorre RJ, Worsley KJ, Evans AC (1996) Modulation of cerebral blood flow in the human auditory cortex during speech: role of motor-to-sensory discharges. Eur J Neurosci 8:2236-2246.

Peeva MG, Guenther FH, Tourville JA, Nieto-Castanon A, Anton JL, Nazarian B, Alario FX (2010). Distinct representations of phonemes, syllables, and supra-syllabic sequences in the speech production network. Neuroimage 50:626-638.

Perkell JS (2010) Movement goals and feedback and feedforward control mechanisms in speech production. J Neurolinguist.

Perkell JS, Cohen MH, Svirsky MA, Matthies ML, Garabieta I, Jackson MT (1992) Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. J Acoust Soc Am 92:3078-3096.

Perkell JS, Lane H, Denny M, Matthies ML, Tiede M, Zandipour M, Vick J, Burton E (2007) Time course of speech changes in response to unanticipated short-term changes in hearing state. Journal of the Acoustical Society of America 121:2296-2311.

Perkell J, Matthies M, Lane H, Guenther F, Wilhelms-Tricarico R, Wozniak J, Guiod P (1997) Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. Speech Commun 22:227-250.

Perkell JS, Matthies ML, Svirsky MA, Jordan MI (1993) Trading Relations between Tongue-Body Raising and Lip Rounding in Production of the Vowel-U - a Pilot Motor Equivalence Study. J Acoust Soc Am 93:2948-2961.

Perkell J, Numa W, Vick J, Lane H, Balkany T, Gould J (2001) Language-specific, hearing-related changes in vowel spaces: A preliminary study of English- and Spanish-speaking cochlear implant users. Ear Hearing 22:461-470.

Perrier P, Ostry DJ, Laboissiere R (1996). The equilibrium point hypothesis and its application to speech motor control. J Speech Hear Res 39:365-378.

Peters RW (1954). The effect of changes in sidetone delay and level upon rate of oral reading in normal speakers. J Speech Hear Disord 20:371-375.

Platt JR (1964) Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. Science 146:347-353.

Porfert AR, Rosenfield DB (1978) Prevalence of stuttering. Journal of Neurology, Neurosurgery, and Psychiatry 41:954-956.

Postma A, Kolk H (1993) The covert repair hypothesis: Prearticulatory repair processes in normal and stuttered disfluencies. J Speech Hear Disord 36:472-487.

Prosek RA, Montgomery AA, Walden BE, Hawkins DB (1987) Formant frequencies of stuttered and fluent vowels. J Speech Hear Res 30:301-305.

Purcell DW, Munhall KG (2006a) Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. Journal of the Acoustical Society of America 120:966-977.

Purcell DW, Munhall KG (2006b) Compensation following real-time manipulation of formants in isolated vowels. Journal of the Acoustical Society of America 119:2288-2297.

Quatieri TF (2001) Discrete-time Speech and Signal Processing: Principles and Practice. Upper Saddle River, NJ: Prentice-Hall.

Rhodes BJ, Bullock D, Verwey WB, Averbeck BB, Page MP (2004) Learning and production of movement sequences: behavioral, neurophysiological, and modeling perspectives. Hum Mov Sci 23:699-746.

Riaz N, Steinberg S, Ahmed J, Pluzhnikov A, Riazuddin S, Cox NJ, Drayna D (2005) Genomewide significant linkage to stuttering on chromosome 12. Am J Human Genet 76:647-651.

Riley GD (2008) SSI-4: Stuttering Severity Instrument: PROED.

Ringel RL, Minifie FD (1966) Protensity estimates of stutterers and nonstutterers. J Speech Hear Res 9:289-296.

Ringel RL, Steer MD (1963).Some Effects of Tactile and Auditory Alterations on Speech Output. J Speech Hear Res 6:369-378.

Salmelin R, Schnitzler A, Schmitz F, Jancke L, Witte OW, Freund H-J (1998) Functional organization of the auditory cortex is different in stutterers and fluent speakers. NeuroReport 9:2225-2229.

Saltzman ELM, Munhall KG (1989). A dynamical approach to gestural patterning in speech production. Ecol Psychol 1:333-382.

Saltzman E, Nam H, Goldstein L, Byrd D (2006) The distinctions between state, parameter and graph dynamics in sensorimotor control and coordination. In: M.L. Latash and F. Lestienne (Eds), Progress in Motor Control: Motor Control and Learning over the Life Span, pp. 63-73.

Savariaux C, Perrier P, Orliaguet JP (1995) Compensation Strategies for the Perturbation of the Rounded Vowel [U] Using a Lip Tube - a Study of the Control Space in Speech Production. J Acoust Soc Am 98:2428-2442.

Shadmehr R, Mussaivaldi FA (1994) Adaptive Representation of Dynamics during Learning of a Motor Task. J Neurosci 14:3208-3224.

Siegel GM, Pick HL, Jr. (1974) Auditory feedback in the regulation of voice. J Acoust Soc Am 56:1618-1624.

Shaiman S, Gracco VL (2002) Task-specific sensorimotor interactions in speech production. Exp Brain Res 146:411-418.

Shattuck-Hufnagel S (1987) The role of word onset consonants in speech production planning: new evidence from speech error patterns. In: Motor and Sensory Processes of Language, pp 17-53 Hillsdale, NJ: Laurence Erlbaum.

Silverman FH, Williams DE (1967) Loci of disfluencies in the speech of stutterers. Percept Mot Skills 24:1085-1086.

Soderberg G (1966) The relations of stuttering to word length and word frequency. J Speech Hear Res 9:584-589.

Sommer M, Koch MA, Paulus W, Weiller C, Buchel C (2002) Disconnection of speech-relevant brain areas in persistent developmental stuttering. Lancet 360:380-383.

Song LP, Peng DL, Jin Z, Yao L, Ning N, Guo XJ, Zhang T (2007) [Gray matter abnormalities in developmental stuttering determined with voxel-based morphometry]. (In Chinese) Zhonghua Yi Xue Za Zhi (Chinese Journal of Medicine) 87:2884-2888.

Spencer KA, Slocomb DL (2007) The neural basis of ataxic dysarthria. Cerebellum 6:58-65.

Stager SV, Denman DW, Ludlow CL (1997) Modifications in aerodynamic variables by persons who stutter under fluency-evoking conditions. J Speech Lang Hear R 40:832-847.

Stager SV, Jeffries KJ, Braun AR (2003) Common features of fluency-evoking conditions studied in stuttering subjects and controls: an (H2OPET)-O-15 study. J Fluency Disord 28:319-336.

Stark R, Levitt H (1974). Prosodic feature reception and production in deaf children (abstract). J Acoust Soc Am 55:S563(A).

Stark RE, Pierce BR (1970) The effects of delayed auditory feedback on a speech-related task in stutterers. J Speech Hear Res 13:245-253.

Starkweather CW (1987) Fluency and Stuttering. Englewood Cliffs, NJ: Prentice-Hall.

Stephen SCG, Haggard MP (1980) Acoustic properties of masking/delayed auditory feedback in the fluency of stutterers and controls. J Speech Hear Res 23:527-538.

Stevens KN (1998) Acoustic phonetics. Cambridge, Mass.: MIT Press.

Suresh R, Ambrose NG, Roe C, Pluzhnikov A, Wittke-Thompson JK, Ng MC, Wu X, Cook EH, Lundstrom C, Garsten M (2006) New complexities in the genetics 904 of stuttering: significant sex-specific linkage signals. Am J Human Genet 78:554-563.

Sussman HM, Smith KU (1971). Jaw movements under delayed auditory feedback. J Acoust Soc Am 50:685-691.

Sutton C, Chase R (1961) White noise and stuttering. J Speech Hear Disord 4:72.

Tourville JA, Reilly KJ, Guenther FH (2008).Neural mechanisms underlying auditory feedback control of speech. Neuroimage 39:1429-1443.

Toyomura A, Koyama S, Miyamaoto T, Terao A, Omori T, Murohashi H, Kuriki S (2007) Neural correlates of auditory feedback control in human. Neuroscience 146:499-503.

Tremblay S, Shiller DM, Ostry DJ (2003) Somatosensory basis of speech production. Nature 423:866-869.

Tremblay S, Houle G, Ostry DJ (2008) Specificity of speech motor learning. J Neurosci 28:2426-2434.

Tye-Murray N, Kirk KI (1993) Vowel and diphthong production by young users of cochlear implants and the relationship between the phonetic level evaluation and spontaneous speech. Journal of Speech and Hearing Research 36:488-502.

Tye-Murray N, Spencer L (1995) Acquisition of speech by children who have prolonged cochlear implant experience. J Speech Lang Hear Res 38:327-337.

Villacorta VM, Perkell JS, Guenther FH (2007) Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. J Acoust Soc Am 122:2306-2319.

Van Riper C (1982) The Nature of Stuttering. Prospect Heights, IL: Waveland Press, Inc.

Van Summers W, Pisoni DB, Bernacki RIP, Stokes MA (1988) Effects of noise on speech production: Acoustic and perceptual analyses. J Acoust Soc Am 84:917-928.

Ventura MI, Nagarajan SS, Houde JF (2009) Speech target modulates speaking induced suppression in auditory cortex. BMC Neuroscience 10:58.

Villacorta VM, Perkell JS, Guenther FH (2007) Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. J Acoust Soc Am 122:2306-2319.

Waldstein RS (1990). Effects of Postlingual Deafness on Speech Production - Implications for the Role of Auditory-Feedback. Journal of the Acoustical Society of America 88:2099-2114.

Waltz RA, Morales JL, Nocedal J, Orban D (2006) An interior alogirhm for nonlinear optimization that combines line search and trust region steps. Mathematical Programming 107:391-408.

Wang W, Lan J, Song MS, Pan CH, Zhuang GQ, Wang YX, Ma WZ, Chu QY, Lai QX, Xu F, Li YL, Liu LX (2009) Association between dopaminergic genes (SLC6A3 and DRD2) and stuttering among Han Chinese. J Hum Genet 54:457-460.

Watkins KE, Smith SM, Davis S, Howell P (2008) Structural and functional abnormalities of the motor system in developmental stuttering. Brain 131:50-59.

Webster RL, Schumacher SJ, Lubker BB (1970) Changes in stuttering frequency as a function of various intervals of delayed auditory feedback. J Abnorm Psychol 75:45-49.

Wieneke GH, Eijken E, Janssen P, Brutten GJ (2001) Durational variability in the fluent speech of stutterers and nonstutterers. J Fluency Disord 26:43-53.

Wildgruber D, Ackermann H, Grodd W (2001) Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: effects of syllable repetition rate evaluated by fMRI. Neuroimage 13:101-109.

Wittke-Thompson JK, Ambrose N, Yairi E, Roe C, Cook EH, Ober C, Cox NJ (2007) Genetic studies of stuttering in a founder population. J Fluency Disord 32:33-50.

Wolpert DM, Ghahramani Z, Jordan MI (1995) An internal model for sensorimotor integration. Science 269:1880-1882.

Wolpert DM, Kawato M (1998) Multiple paired forward and inverse models for motor control. Neural Netw 11:1317-1329.

Wolpert DM, Miall RC, Kawato M (1998) Internal models in the cerebellum. Trends in Cognitive Sciences 2:338-347.

Xia K, Espy-Wilson CY (2000) A new strategy of formant tracking based on dynamic programming. In: Sixth International Conf on Spoken Language Processing (IC-SLP2000) Beijing, China.

Xu Y, Larson CR, Bauer JJ, Hain TC (2004) Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. Journal of the Acoustical Society of America 116:1168-1178.

Yairi E, Ambrose N (2005) Early Childhood Stutering: For Clinicians by Clinicians. Austin, Texas: Pro-Ed.

Zimmermann G, Brown C, Kelso JAS, Hurtig R, Forrest K (1988) The association between acoustic and articulatory events in a delayed auditory-feedback paradigm. Journal of Phonetics 16:437-451.