Linguistic constraints on compensation for altered auditory feedback

by

Shira Eden Katseff

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Linguistics

in the

Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor Keith Johnson, Chair
Professor John Houde
Professor Susanne Gahl
Professor Dan Klein

Fall 2010

Linguistic constraints on compensation for altered auditory feedback

by Shira Eden Katseff

Abstract

Linguistic constraints on compensation for altered auditory feedback

by

Shira Eden Katseff

Doctor of Philosophy in Linguistics

University of California, Berkeley

Professor Keith Johnson, Chair

With great effort, adults can try to produce new language sounds, with varying degrees of success, yet these same adults automatically and routinely adjust their speech production to accommodate their environments and interlocutors.

This adjustment process allows talkers to be heard above traffic, speak clearly to foreigners, and speak intelligibly with pens in their mouths. All of these actions require *monitoring auditory feedback*, listening to one's outgoing speech and ensuring that it is error-free. This dissertation investigates the mechanism that allows this automatic maintenance to proceed.

The four experiments described here use a device that systematically alters auditory feedback in real time. Subjects in these experiments hear a slightly distorted version of their own voices, in which one or two of their vowel formants have been shifted. Talkers automatically adjust to these manipulations by *opposing* the shift in auditory feedback; if they hear their vowels with an artificially high first formant, they start producing vowels with lower first formants. That is, *talkers change their speech production to compensate for the shifted auditory feedback*.

The first experiment demonstrates that talkers do not change their speech in the way one might naively expect; they do not produce a vowel change exactly inverse to what they hear. The remaining three experiments attempt to characterize what subjects are optimizing when they compensate for shifts in auditory feedback, with particular attention to whether language-specific sound and word patterns constrain automatic speech processing.

In particular, they ask whether low-level compensation for altered auditory feedback is influenced by top-down information from a talker's phonological or lexical inventory, or by acoustic familiarity.

Experiments 3 and 4 find that talkers seek to avoid confusability, compensating less in regions of vowel space with multiple competing vowels, and choose compensation routes that run through acoustically familiar regions of vowel space. Experiment 2 fails to find consistent evidence of influence from the lexical inventory. Together, these experiments demonstrate that even automatic speech processes that operate

on low-level auditory information are influenced by high-level knowledge about one's phonological inventory and social context. Results of these experiments are used to expand models of speech motor control and to unify them with psycholinguistic models of speech production.

Professor Keith Johnson
Dissertation Committee Chair

To toil.

Contents

List of Figures iii

List of Tables x

1 Introduction 1

2 Background 5

3 Methods 38

4 Experiment 1: Word avoidance in compensation for auditory feedback shift 59

5 Experiment 2: Somatosensory and phonemic influences on compensation for shifts in auditory feedback 82

6 Experiment 3: Modeling individual variation in compensation 111

7 General Discussion 140

References 148

A 161

B 165

List of Figures

2.1 Vowel formant transitions in American English. Reprinted with permission from Delattre, Liberman, & Cooper, 1955. Copyright 1955, Acoustical Society of America. 16
2.2 Current schematic of Directions into Velocities of Articulators (DIVA) model. From Tourville & Guenther, 2010. 27
2.3 Current schematic of the State Feedback Control (SFC) model. From Houde & Nagarajan, under review. 29
2.4 Two studies of Standard American English vowel spaces, as reinterpreted by Hagiwara (1997). Men's vowels are in open squares and women's vowels are in filled circles. Left: Peterson & Barney's General American English Vowels. Reprinted with permission. Copyright 1997, Acoustical Society of America. Right: Hillenbrand et al.'s (1995) Standard American English vowels. Reprinted with permission. Copyright 1995, Acoustical Society of America. 35
2.5 Two studies of California English vowel spaces. Left: Data collected by Hagiwara (1997). Men's vowels are in open circles, and women's vowels are in filled circles. Right: Formants from Clopper et al. (2005) Western male vowels, for comparison (all tokens are shown with an ellipse, hand-drawn by the authors, surrounding each cluster). Reprinted with permission. Copyright 2005, Acoustical Society of America. 36
2.6 Two studies of California English vowel spaces. Left: Normalized formants from Hagiwara 1997's male and female California English speakers. Reprinted with permission. Copyright 1997, Acoustical Society of America. Male speaker averages are in blue and female speaker averages are in orange. Right: Vowel formants collected by Hall-Lew (2009). All tokens are graphed on normalized fomant axes. 37

3.1 Schematic of Experimental Setup. Subjects speak into the microphone portion of a headset. Their speech is analyzed, then re-synthesized (and shifted, if necessary) and fed back into the headset's earphones. 39

3.2 Fourier transform of mid-vowel frame from 'food'. The smoothed spectral envelope is overlaid on the spectrum. As marked, formants are peaks in the smoothed spectrum. Formant shifting simply requires changing the value of the formant of interest. The peaks calculated in this particular spectral slice are: F1 = 301 Hz; F2 = 1924 Hz; F3 = 2367 Hz. 40
3.3 Fourier transform of mid-vowel frame from 'food' with direction of shift marked. 41
3.4 Outbuffer from mid-vowel frame of 'food'. The second formant has been shifted by 300 Hz such that it now lies on top of F3. Here, one frame from the shifted word (at 0.57 sec.) is shown. 42
3.5 Top: Change in F1 feedback and F1 production in /ɛ/ over the course of the experiment for a representative subject. Gray filled circles mark the F1 values that the subject heard at each trial. Open circles mark the F1 that the subject produced at each trial. Thus, each gray circle/open circle pair represents one trial. Bottom: Change in F1 feedback and F1 production in /ɛ/ over the course of the experiment, averaged across subjects from Purcell & Munhall, 2006. Reprinted with permission. Copyright 2006, Acoustical Society of America. Open circles indicate the F1 values that the subject heard at each trial, and black filled circles mark the F1 that the subject produced at each trial. 44
3.6 Top: Vowel trajectories for a speaker of California English. Each line segment represents one vowel produced by this speaker. Formants at vowel onset are at the unlabeled end of the line segment, and formants at vowel offset are at the labeled end of the line segment. Bottom: Trajectories of midwestern American English vowels, as reported by Hillenbrand and Nearey (1999). 49
3.7 Change in feedback over the course of each experiment. There were 360 trials in each experiment. These 360 trials were composed of 6 regions of equal formant alteration ("stairs") connected by ramps of slowly increasing or decreasing feedback alteration. 50
3.8 F1 from the /ɛ/ in 'head' over the course of the experiment (one typical subject shown here). Open circles mark F1 from the vowels that the subject produced at each trial, and gray filled circles show the altered F1 heard by the subject at each trial. Each gray circle/open circle pair represents one trial. 51
3.9 A typical subject's baseline region for the /ɛ/ in 'head'. Open circles mark the vowel formants extracted from the 360 vowels produced during the control condition, and the solid black line is the convex hull surrounding them. Gray circles mark the vowel formants produced during unaltered trials of the F1 shift experiment, and the smaller, gray convex hull outlines them. 52

3.10 Percent compensation, averaged across subjects, at each of the five formant shift plateaus, as estimated from a linear mixed effects model. Error bars mark 95% confidence intervals for each plateau. 54
3.11 Productions of /ɛ/ during Experiment 1 plotted in F1-F2 Bark space against productions of /ɛ/ during control trials. Results for a typical subject are shown. For small feedback shifts, the light gray shape (formants heard as a result of the feedback shift) falls entirely within the dashed shape (the subject's baseline range), indicating that the vowels that the subject heard were all within his baseline region and that compensation was complete. As the amount of feedback shift increases (the dark solid shape), compensation is less and less complete. 55
3.12 Raw compensation in Hertz, averaged across subjects, at each of the five formant shift steps. Error bars mark 95% confidence intervals for each plateau. 56

4.1 Two points in the speech motor control system where the lexicon might be accessed. 60
4.2 Illustration of the two experimental conditions. In Condition 1, the subject produced the nonword 'deg' as it sounded increasingly like the word 'dig'. In Condition 2, the subject produced the nonword 'teg' as it sounded more and more like the nonword /tɪg/. 62
4.3 /ɛ/ formant change over the course of the 'deg' condition of the experiment. All stimuli were 'deg'. The top graph shows F1 at each trial, and the bottom graph shows F2 at each trial. Formants produced by the subjects are shown in black dots and formants heard by the subjects are shown in blue dots. Each black dot - blue dot pair represents the average of what all subjects said and heard during one trial. Each subject's measurements were scaled by subtracting the median baseline formants. Notice that subjects compensate by opposing the formant shift in both F1 and F2. 64
4.4 /ɛ/ formant change over the course of the 'teg' condition of the experiment. All stimuli were 'teg'. As above, the top graph shows F1, and the bottom graph shows F2 at each trial; blue dots indicate the formants that the subject heard, and black dots indicate the formants that the subjects produced. All formant measurements were scaled by subtracting the median of the first 15 (no-shift) trials. 65

4.5 Calculations used for wedge plots. Compensation magnitude is the Euclidean distance between the formants produced during the first 15 (non-shift) trials and the formants produced during the last 40 trials of the maximum formant shift (measured in Bark). The gray dotted line shows the compensation expected if the subject opposed the formant shift directly and completely. The compensation angle is the angle between the expected end formants and the average formants actually produced during the last 40 maximum shift trials. 68
4.6 (a) Summary of results for both experimental conditions. Wedges display the middle 50% of formants produced across subjects during the last 40 trials with maximum formant shift. Formants were normalized with the magnitude-angle method described above. The darker wedge shows data from the 'deg' condition, and the lighter wedge shows data from the 'teg' condition. 69
4.7 Individual results for 'deg' and 'teg' experimental conditions. Wedges display the formant difference between the first 15 unaltered trials and the last 40 trials from the maximum shift condition. The inner and outer bounds of the wedge mark the 25th and 75th magnitude measurement percentiles, and the lower and upper radii mark the 25th and the 75th percentile angles (see text for a more complete description). The darker wedge shows 'deg' data, and the lighter wedge shows 'teg' data. Only subject 42 shows a difference between conditions. 70
4.8 (a) Summary of results for both experimental conditions. Wedges display the formant difference between the first 15 unaltered trials and the last 40 words from the maximum shift condition. The inner and outer bounds of the wedge mark the 2.5th and 97.5th magnitude measurement percentiles, and the lower and upper radii mark the 2.5th and the 97.5th percentile angles (see text for a more complete description of magnitude and angle calculations). The darker wedge shows data from the 'deg' condition, and the lighter wedge shows data from the 'teg' condition. 71
4.9 /ɛ/ formant change during the the Communicative deg/teg task. Averages for 'deg' stimuli are shown. As above, the top graph shows F1, and the bottom graph shows F2 at each trial; blue dots indicate the formants that the subjects heard, on average, and black dots indicate the formants that the subjects produced, on average. All formant measurements were scaled by subtracting the median of the first 15 (no-shift) trials. 74

4.10 /ɛ/ formant change during the the Communicative deg/teg task. Averages for 'teg' stimuli are shown. As above, the top graph shows F1, and the bottom graph shows F2 at each trial; blue dots indicate the formants that the subjects heard, and black dots indicate the formants that the subjects produced. All formant measurements were scaled by subtracting the median of the first 15 (no-shift) trials. 75
4.11 /ɛ/ formant change during the the Communicative deg/teg task. The average interquartile range (25% - 75%) for 'deg' and 'teg' stimuli are shown. Across subjects, there is no difference in compensation magnitude for the two stimuli, and a small difference in the compensation angles for the two stimuli. Neither is at all confusable with adjacent vowel /ɪ/. 76
4.12 Individual results for Communicative 'deg' / 'teg' task. Wedges display the formant difference between the first 15 unaltered trials and the last 25 words from the maximum shift condition. Wedges mark the 25th and 75th magnitude and angle percentiles (see text for a more complete description of calculations). The darker wedge shows data from the 'deg' condition, and the lighter wedge shows data from the 'teg' condition. 77
4.13 (Left) Histogram of 10,000 'teg' F1 samples minus 10,000 'deg' F1 samples from MCMC. (Right) Histogram of 10,000 'teg' F2 samples minus 10,000 'deg' F2 samples from MCMC. Neither histogram supports the lexical interference hypothesis. 79

5.1 Two points in the speech motor control system where the phoneme inventory might be accessed. Contribution from the phoneme inventory to feedback makes talkers especially wary of *hearing* an out-of-category vowel. Contributions of the phoneme inventory to the adjustment process would make talkers especially wary of *saying* an out-of-category vowel. 86
5.2 Mean formant frequencies of vowels for male speakers of California English in this experiment. 89
5.3 (Left) Change in feedback over the course of the 200-trial (/ʌ/ and /u/) experiments. Both experiments began with a 15-trial period of veridical feedback, followed by a 65-trial ramp up to the maximum feedback shift. For the last 30 trials, feedback was unshifted. (Right) There were 360 trials in the /ɛ/ experiment. These 360 trials were composed of 6 regions of equal formant alteration (plateaus) connected by ramps of slowly increasing feedback alteration. 92

5.4 Response to /ɛ/ F2 feedback shift. Blue dots show the formants that subjects heard at each trial, and the black dots show the formants that they produced at each trial. Each dot represents an average across all /ɛ/ subjects. 93
5.5 Response to /ʌ/ F2 feedback shift. The x-axis marks the trial number. Blue dots mark the normalized formant that the talkers head, on average, at each trial. Black dots show the formants they produced at each trial, on average. The top graph shows F1 over the course of the experiment, and the bottom graph shows how subjects compensate for a shift in /ʌ/ F2. See text for normalization procedure. 95
5.6 Response to /u/ F2 feedback shift. 96
5.7 Wedges display the 25th and 75th percentiles of magnitude and angle response to F2 feedback shift in /ɛ/, /ʌ/, and /u/. (Top left) The middle 50% of formants produced across subjects in response to a 250Hz shift in F2 /ɛ/ feedback. (Top right) The middle 50% of formants produced across subjects during the last 40 trials with 400Hz shift in the F2 of /ʌ/. (Bottom) The middle 50% of formants produced across subjects during the last 40 trials with 300Hz shift in the F2 of /u/. Formants were normalized with the magnitude-angle method described above. 98
5.8 Green "x" shows the mean formants heard during the maximum F2 feedback shift in /ɛ/ (top left); /ʌ/ (top right); /u/ (bottom). Formants were normalized with the magnitude-angle method described above. Wedges display the 25th and 75th percentiles of compensation magnitudes and angles during these trials. 100
5.9 Amended model of speech motor control. 102
5.10 Response to 30 and 90Hz /ɛ/ F1 feedback shift. 105
5.11 Response to 30 and 90Hz /ɛ/ F1 feedback shift. Leftmost graph: Normalized formants produced under no formant feedback shift. Center graph: Normalized formants produced with a 30Hz feedback shift in F1. Rightmost graph: Normalized formants produced with a 90Hz feedback shift in F1. 105
5.12 Wedges display the 25th and 75th percentiles of magnitude and angle response to the maximum (90Hz) F1 feedback shift in /ɛ/ across subjects. Formants were normalized with the magnitude-angle method described earlier. 107

6.1 Vowel densities for 2 females and 2 males, taken from the Buckeye corpus. s16 and s18: old female; s19: old male; s33: young male. . . . 114
6.2 An older female's /æ/ and /ɛ/ vowel densities, respectively. Notice that there is a great deal of overlap. 115

6.3 Vowel density map of casual speech from a male native English speaker, created from (Left) 4654 vowels extracted from a 30-minute mock interview, and (Right) 360 vowels in citation form CVC words. Vowels are not labeled in either map. Density is depicted in heat colors. The lighter the square, the more vowel formants were produced in that square. Vowels in black squares were never produced. 116
6.4 Histogram of trial-to-trial change in formants for subject, pooled across 5 experiments. Height of bars indicates number of trials with the indicated amount of formant change. 121

7.1 Unified model of speech production. 144

A.1 Response to 30 Hz /ɛ/ F2 feedback shift. 163
A.2 Response to 90 Hz /ɛ/ F2 feedback shift. 164

List of Tables

4.1 Summary of experimental manipulation. At the beginning of the experiment, subjects say and hear themselves say 'deg' or 'teg'. If subjects do not compensate at all, they hear themselves saying /tɪg/ or 'dig' at the maximum shift. If subjects compensate completely, they produce (and receive somatosensory feedback reflecting) 'tag' or /dæg/. 61
4.2 Summary of mixed-effect linear regression models comparing the last 40 'deg' trials with maximum shift to the corresponding 'teg' trials for each of the 5 subjects who participated in both the 'deg' and 'teg' conditions. There is no effect of condition in F1, but a significant effect of condition in F2. 66
4.3 Summary of experimental manipulation. At the beginning of the experiment, subjects will say and hear themselves say 'deg' or 'teg'. If subjects do not compensate at all, they will hear themselves saying 'tag' or /dæg/ at the maximum shift. If subjects compensate completely, they will produce (and receive somatosensory feedback reflecting) /tɪg/ or 'dig'. 73
4.4 Summary of mixed-effect linear regression models comparing the 'deg' trials with maximum shift to the 'teg' trials with maximum for all subjects in the communicative version of the deg/teg experiment, with 'deg' as baseline. Compensation for the two stimulus words are significantly different in both formants, though **not** in the expected direction. 78

5.1 Proxy for salience of somatosensory feedback in three English vowels. 88
5.2 Important characteristics of acoustic and somatosensory targets for three English vowels. 89
5.3 Predicted influence of acoustic and somatosensory information on compensation across the vowel space. 90
5.4 Experimental design, /ʌ/ and /u/ conditions. 91
5.5 Summary of mixed-effect linear regression model estimating F1 and F2 compensation for /ɛ/ at maximum shift with subject as a random effect. Compensation is significantly different from 0 in both F1 and F2. 94

5.6 Summary of mixed-effect linear regression model estimating F1 and F2 compensation for /ʌ/ at maximum shift with subject as a random effect. Compensation is not significantly different from 0 in F1 and marginally different from 0 in F2. 94
5.7 Summary of mixed-effect linear regression model estimating F1 and F2 compensation for /u/ at maximum shift with subject as a random effect. Compensation is significantly different from 0 in F2, but not F1. 96
5.8 Average magnitudes and angles of compensation across subjects for the three altered vowels. 99
5.9 Predicted and actual effects of acoustic and somatosensory information on compensation across the vowel space. 101
5.10 Design of small F1 shift experiment. 104
5.11 Kolmogorov-Smirnov tests for 30 Hz and 90 Hz shifts in the 'head' vowel. 106

6.1 Correlations between properties of experiment and background vowel space from mock interview. 119
6.2 Correlations between properties of experiment and background vowel space from citation form words. 119
6.3 Predictors in linear regression predicting F1 in casual speech. 124
6.4 Predictors in linear regression predicting F2 in casual speech. 125
6.5 Summary of regressor coefficients in model predicting F1 in casual speech from a mock interview. Factors in boldface were significant as main effects in the pooled experiment analysis. Cells marked with * are significant with Pr(>| t |) between 0.01 and 0.05. Cells marked with ** are significant with Pr(>| t |) between 0.001 and 0.01. Cells marked with *** are significant with Pr(>| t |) <0.001. 125
6.6 Summary of regressor coefficients in model predicting F2 in casual speech from a mock interview. Factors in boldface were significant as main effects in the pooled experiment analysis. Cells marked with * are significant with Pr(>| t |) between 0.01 and 0.05. Cells marked with ** are significant with Pr(>| t |) between 0.001 and 0.01. Cells marked with *** are significant with Pr(>| t |) <0.001. 126
6.7 β coefficients of vowel density, completeness of compensation, and the probability of hearing the intended vowel that are associated with the maximum likelihood of the formant trajectory data. Data from each experiment was analyzed separately. Density information was extracted from casual speech. 131
6.8 Results of the optimization analysis, data pooled across all 5 experiments. β coefficients of vowel density, completeness of compensation, and the probability of hearing the intended vowel that are associated with the maximum likelihood of the formant trajectory data. Density information was extracted from casual speech. 131

6.9 Results of the optimization analysis using data from all experiments *except* the one indicated in the column heading. β coefficients of vowel density, completeness of compensation, and the probability of hearing the intended vowel that are associated with the maximum likelihood of the formant trajectory data are listed. Density information was extracted from casual speech. 132
6.10 Predictors in linear regression predicting F1 in citation form speech. . 134
6.11 Predictors in linear regression predicting F2 in citation form speech. . 135
6.12 Summary of regressor coefficients in model predicting F1 in citation speech. Factors in boldface were significant as main effects in the pooled experiment analysis. Cells marked with * are significant with Pr(>| t |) between 0.01 and 0.05. Cells marked with ** are significant with Pr(>| t |) between 0.001 and 0.01. Cells marked with *** are significant with Pr(>| t |) <0.001. 135
6.13 Summary of regressor coefficients in model predicting F2 in citation speech. Factors in boldface were significant as main effects in the pooled experiment analysis. Cells marked with * are significant with Pr(>| t |) between 0.01 and 0.05. Cells marked with ** are significant with Pr(>| t |) between 0.001 and 0.01. Cells marked with *** are significant with Pr(>| t |) <0.001. 136
6.14 Predictors in optimization analysis finding the weights associated with the maximum likelihood of the data and subject-specific information about vowel density, completeness of compensation, and the probability of hearing the intended vowel. 137

B.1 Predictors in linear regression predicting F1 in citation vowels. 165
B.2 Predictors in linear regression predicting F1 in citation vowels. 166
B.3 Predictors in linear regression predicting F1 in citation vowels. 166
B.4 Predictors in linear regression predicting F1 in citation vowels. 166
B.5 Predictors in linear regression predicting F1 in citation vowels. 166
B.6 Predictors in linear regression predicting F1 in citation vowels. 166
B.7 Predictors in linear regression predicting F1 in citation vowels. 167
B.8 Predictors in linear regression predicting F1 in citation vowels. 167
B.9 Predictors in linear regression predicting F1 in citation vowels. 167
B.10 Predictors in linear regression predicting F1 in citation vowels. 167
B.11 Predictors in linear regression predicting F1 in casual speech. 168
B.12 Predictors in linear regression predicting F1 in casual speech. 168
B.13 Predictors in linear regression predicting F1 in casual speech. 168
B.14 Predictors in linear regression predicting F1 in casual speech. 169
B.15 Predictors in linear regression predicting F1 in casual speech. 169
B.16 Predictors in linear regression predicting F2 in casual speech. 169
B.17 Predictors in linear regression predicting F2 in casual speech. 169

B.18 Predictors in linear regression predicting F2 in casual speech. 169
B.19 Predictors in linear regression predicting F2 in casual speech. 170
B.20 Predictors in linear regression predicting F2 in casual speech. 170

Acknowledgments

Thank you, mentors, colleagues, friends, and family.

Completing this research project was challenging, but it would have been impossible without my well-rounded advisor team. Keith Johnson, thanks for supporting my many oddball ideas, including the one that grew into this dissertation. Your advice and bookshelf have come in handy more times than I can count. John Houde, you have finally convinced me that the brain is worth studying. Thanks for your generosity with your equipment and software, and for treating me as a member of your lab. Susanne Gahl, you have made me a more careful scientist. Your honest critiques have been great for keeping my questions and analyses in line. Dan Klein, thanks for sharing your perspective on my work and serving on my committee.

This work would also have been impossible without funding. To The National Science Foundation, the Abigail Hodgen Publication Fund, and the Acoustical Society of America: much obliged.

This project you're about to read about required a dedicated computer, which Ronald Sprouse kindly helped me install, maintain, and keep in an already-crowded sound booth. Thanks for rescuing me from myriad hardware and software jams! It also required many semesters' worth of painstaking and sometimes tedious analysis. I am hugely grateful to Matt LaCoste and Svetlin Dimov, whose help with running subjects and preprocessing vowel data saved my time and sanity. My subjects also have my gratitude for letting me pick apart their very cool Californian vowels.

Developing, executing, and writing about this stuff was much easier thanks to conversation and comments from many friends and colleagues. I appreciate you, phonology lab, especially Melinda, Reiko, and Molly; officemates Yao and Russell; quasi-officemate Will; members of John Houde's lab; and Becca, the best comotivator in town.

My friends and family get substantial credit for keeping me out of the lab. Irena, thanks for the hikes, dinners, and constant encouragement! Lauren and Jenny, thanks for reminding me of home. Thank you, mom, Eli, dad, Bonnie, and grandma for caring about me as I went off to seek my fortune. Thanks to Jared for making me laugh about my papers and posters, and to Adiya, whose knitting skills kept my iPod and hands warm. Special thanks go to Richard, my fellow researcher, writer, half marathon runner, and explorer. If I see anything clearly, it's your doing.

I owe Nina Dronkers at the Martinez VA a huge thank you for her mentorship, support, and, most of all, showing me the value of linguistics outside of the classroom. The rest of the aphasia lab and affiliates deserve a shout-out, too: Juliana, Analia, Carl, Lisa, And, Adeen, Janet, Andrea. Thanks for the good conversation and lunchtime company!

Finally, thanks to Jen Hay, Megan McAuliffe, and the rest of the New Zealand Institute for Language, Brain, and Behaviour, for inviting me to start my next chapter. Can't wait!

Chapter 1

Introduction

Our vocal tracts change substantially between childhood and adulthood, yet throughout this period we are able to speak and make ourselves understood. Changing our speech to deal with physiological surprises requires *self-monitoring*, listening to our incoming speech and squaring it with the speech we expect to hear. The process that allows our voices to change slowly over the course of development is poorly understood, and currently there is considerable disagreement over even how we decide what we expect to hear and how we compare our expectations with incoming feedback. Understanding this process is essential to understanding how speech perception and production operate because speech plans are adjusted not only during development, but every day as our speech calibrates to noisy streets and accented talkers.

This dissertation brings together evidence from linguistics, engineering, and neuroscience to gain a better understanding of how speech is planned, and how speech plans are adjusted. It does so by presenting four experiments that perturb the speech motor control system, which, in part, adjusts speech plans to accommodate differences between incoming speech feedback and expectations about that feedback. The experiments in this dissertation create an artificial mismatch between observed and expected feedback, and measure changes in produced speech that result from them. From this information, I make inferences about the specificity of expected feedback, and about constraints on compensation for unexpected feedback. For example, when adjusting one's speech, one can seek to maintain some universal set of acoustic relationships or muscular synergies, or, for each phoneme (or syllable) one might strive to maintain a task-specific set of acoustic or articulatory properties.

The key contribution of this dissertation is its claim that compensation for altered auditory feedback is not simply a motor process. The experiments in the dissertation show instead that compensation is influenced by higher-level language-specific information. Specifically, incomplete compensation is shown to be partly due to **top-down influence from the phonological inventory** and **a preference for commonly-produced vowels**. Experiments failed to show consistent top-down influence from the lexical inventory.

Approach

The experiments perturbing the speech motor control system did so by altering auditory feedback. The approach is similar to that taken in Houde, 1997, and the experimental design was similar to Purcell & Munhall, 2006b. Subjects wore a headset wired so that the microphone ran through a computer and back to the headphones; because of this, subjects were able to hear everything they said in real time. The computer was set to either return the subject's voice veridically or, during experiments, to gradually alter one or two vowel formants in the subject's speech. Subjects were recorded while producing single-syllable words over the course of the experiment.

Data

Data was collected, with IRB approval, from 65 male native English speakers living near Berkeley, California. The vast majority were undergraduate students who grew up in California. Subjects came to the phonology lab at the University of California, Berkeley, to be recorded on three or four separate days, depending on the experiment. For one half-hour on each day, they produced monosyllabic CVC words and nonwords while wearing a Feedback Alteration Device. The Feedback Alteration Device recorded both the speech that they produced and the altered speech that they heard. Their voices were analyzed to determine whether their vowel production changed as their vowel feedback was altered.

Challenges

Other studies have measured the response to altered auditory feedback using similar apparatuses and have encountered a number of challenges. First, altering speech in real time is a technological challenge, even with today's computers. In order to alter formants, the device must first be able to find those formants within a few milliseconds. Accurate formant finding in this time frame requires some compromises. One must work with a low sampling rate, allowing frequencies up to just 4KHz to 5.5Hz. Vowel information is well represented in this frequency range, but fricatives are not. Because of this, the quality of the heard speech is somewhat compromised. In addition, the formant alteration device used in this dissertation changes formants by re-synthesizing incoming speech from formant information, along with estimated pitch and spectral energy. As a consequence, accurate re-synthesis depends on not only accurate formant measurement, but also accurate estimation of bandwidth and harmonics. Real speech is complex enough that it is very difficult to generate an accurate facsimile of a person's voice from these few parameters within such a short period of time. For this reason, subjects often reported that their voices sounded somewhat "distorted", even in the absence of feedback alteration.

A second issue is measurement. Measuring a subject's change in production is not straightforward. It requires quantifying vowel production without alteration and

quantifying vowel production with alteration. Doing so requires making a number of choices. For example, it has been observed that subjects respond to a feedback change in a single formant with a production change involving multiple formants. If we choose to acknowledge this and measure production changes in multiple formants, how many formants should we choose, and how do we weight them: are production changes in F1 and F2 more important than production changes in F3 and F4? For this reason, there are no standard methods of analysis for formant alteration data. One contribution of this dissertation is an experimental design (not the only possible design) that is able to capture the variables of interest without causing fatigue in experimental subjects.

A third issue is interpretation. Some models of speech motor control posit that speech is planned syllable by syllable, some phoneme by phoneme, and some using other units entirely. One major model of speech motor control (Houde & Nagarajan, under review) relies on an internal simulated vocal tract to generate expectations, and the other major model generates expectations from speech memory (Guenther, 2003). Neither accesses lexical or semantic representations of the intended message while monitoring ongoing speech. These models are quite different from models of the phonological loop from psychology (Baddeley, 1992), or models of speech production from psycholinguistics (Levelt, Roelofs, & Meyer, 1999), or exemplar models of speech production from linguistics (Johnson, 1997). Care has been taken to construe results as broadly as possible and to bring disparate models together when appropriate, but it remains a great challenge to bridge ideas from these different fields.

Outline

The rest of the dissertation is structured as follows.

Chapter 2 reviews background relevant to the models and methodology used in this dissertation. It reviews models of speech production and speech motor control, as well as literature on vowel targets, learning, and on motor control more generally.

Chapter 3 reviews the methods and equipment used in the experiments in the dissertation. The first half explains how the Feedback Alteration Device works. The second half describes a preliminary experiment showing that, although subjects compensate for altered auditory feedback, compensation is less and less complete as formant shifts increase in size.

Chapter 4 describes an experiment asking whether altered auditory feedback is influenced by a subject's lexical inventory. This particular experiment compares compensation for a minimal pair of nonwords. The nonwords were chosen so that compensating for one would require producing a word, while compensating for the other would not require producing a word. If subjects accessed their lexical inventories when making production decisions, they were expected to compensate less in the condition where compensation would require a subject to produce an unintended word.

Chapter 5 describes an experiment exploring the influence of phonological inventory on compensation for altered auditory feedback by altering feedback in three regions of vowel space. The three vowels differed both in the number of nearby phonemes and in the salience of their somatosensory feedback. If subjects were to access their phonological inventories while adjusting their articulatory plans, they would be expected to compensate less in dense vowel regions (with many nearby phonemes) in order to avoid producing competing vowels. If subjects were sensitive to the salience of their somatosensations when comparing observed to expected feedback, they would be expected to compensate less in vowels with salient somatosensory feedback because in these cases, vowel identity is tied up in the somatosensation of the vowel.

Chapter 6 investigates whether individual variation in compensation is tied to a speaker's articulatory-acoustic habits. In this case study, a subject's performance on a number of altered feedback experiments is analyzed with respect to his background vowel space, as measured (1) by producing 360 citation form CVC words and (2) by producing casual speech during a half-hour mock interview. Several types of analyses are used to predict this subject's vowel formants during the altered feedback experiments from the formants he produced outside of the experiment.

Chapter 7 summarizes and contextualizes these four studies, then uses their results to attempt to unify psycholinguistic models of speech production with models of speech motor control.

Chapter 2

Background

Perception influences production

Language is a two-way system: talkers both hear incoming speech and produce their own speech. This unusual feature has led linguists in a number of subfields to wonder, somewhat independently, about how language perception influences language production (and vice versa). Problems in several of these subfields, whose solution relies on understanding the connection between perception and production, are listed as examples.

How do infants learn their first language? [Acquisition]. Language input determines an infant's first language. Infants exposed to Urdu speak Urdu; infants exposed to French speak French; infants exposed to multiple languages speak all of them. An infant becomes attuned to the sounds of his/her native language over the course of the first year (Werker & Tees, 1984), and a young child's production system slowly grows to produce language sounds that are comprehensible to other native speakers (Kuhl & Meltzoff, 1996; Vorperian & Kent, 2007).

How do adults learn to produce new speech sounds? [Perceptual learning]. Humans tend to pay more attention to dimensions that help in task performance and ignore irrelevant dimensions (Goldstone, 1998). This is as true for tasks learned in adulthood as those learned in infancy. For the task of speaking, perceptual learning is most important when learning to hear a new speech sound in a foreign language. Surprisingly enough, learning to hear the difference between two new speech sounds helps one to say them better. Japanese famously fails to distinguish /ɹ/ from /l/, and speakers of this language have difficulty perceiving the difference between them. Studies of L1 Japanese speakers learning English as an L2 have shown that production after practice with perceiving and hearing the two sounds is superior to an equivalent amount of practice in perceiving the sounds without production practice (e.g. Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997).

Why do talkers pick up phonetic and syntactic characteristics of their interlocutors? [Sociolinguistics]. Recent work shows that talkers passively acquire phonetic

characteristics of their interlocutors as they speak, approaching their vowel formants, in a process called *accommodation* (Pardo, 2006; Babel, 2009). This is also true of laboratory shadowing tasks, in which talkers repeat a sound or word they hear (Goldinger, 1998; Nielsen, 2008), and even of interviewers performing dialect studies (Watt, Llamas, & Johnson, 2010). Talkers are also more likely to use syntactic constructions that they have just heard (Bock, 1986).

Why do talkers produce "extreme" formants when speaking clearly? [Phonetics]. Similarly, some phonetic research investigates why speakers tend to expand their vowel spaces when they wish to increase loudness or intelligibility. So-called *clear speech* is associated with a large improvement in intelligibility over conversational speech (Picheny, Durlach, & Braida, 1985, 1986). It tends to be implemented by producing vowels with extreme formants, increasing the auditory distance between adjacent vowels.

Why are talkers faster to say 'butter' after seeing the word 'bread'? [Psycholinguistics]. The connection between perception and production has been studied in higher-level processing as well. When deciding whether two letter strings are both words or both nonwords, subjects are *primed* to make faster decisions when (a) both items are words, (b) when the items are semantically associated and (c) when the items are frequent (Meyer & Schvaneveldt, 1971). Priming is present not only in commonly associated words, but also in semantically, syntactically, and phonologically related words (Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982).

Why do talkers speak loudly in stadiums and quietly in movie theaters? [Phonetics]. Perception and production interact within single talkers as well. In 1911, Etienne Lombard noticed that talkers in noisy conditions tend to speak more loudly. Talkers will increase loudness automatically when on a noisy street or under a hairdryer, even if their interlocutor is in quieter conditions, and even when they are told that their interlocutor does not need increased volume (Pick, Jr., Siegel, Fox, Garber, & Kearney, 1989). It is possible to explain this behavior as an adjustment made for one's audience rather than for one's own expectations (Lane & Tranel, 1971).

Plan for this chapter

Most of the examples of perception-production influence reviewed thus far show that some aspect of a talker's speech (or speech processing) is changed by hearing speech from a different talker. Understanding these cases is difficult because they require modeling the language processing system in each of the individual talkers as well as the interface between them. The experiments performed in this dissertation distill this problem to the case of a single talker, asking the question: how is one's speech affected by hearing oneself?

What is known so far is that talkers aim for vowel targets as they speak, and that they adjust those targets over time. Long-term target changes are necessary as we grow and our vocal tract changes shape. Short term adjustment is also required,

as when one converses with a foreigner, or speaks with a pipe between one's lips.

In order to understand how the experiments described in this dissertation can illuminate how heard speech is translated to produced speech, it is important to place these experiments in context.

A number of models have been proposed to explain how perception influences production. The experiments in this dissertation can inform these models by intercepting linguistic systems at multiple points in the processing of incoming and outgoing speech. The first section reviews theoretical and experimental evidence for how talkers plan and represent vowels. The second section reviews theories of how vowel representations change over time. The final section discusses how experimental evidence about target change can help to expand or unify models of speech production in psycholinguistics, engineering, and neuroscience.

Units of speech planning

There are two main classes of models that are used to describe speech production. Psycholinguistic models of speech production explain how the intended message is translated into words and phonemes. Speech motor control models explain how a selected speech unit is articulated. Both types of models are starting to converge on syllable-sized planning units.

This convergence stems from a potpourri of psycholinguistic evidence, which was collected with two types of data. One is documentation of observed or experimentally-induced patterns in speech errors. Speech errors appear to differ in character between languages, with some languages tending toward single-phoneme transpositions, and others more prone to consonant cluster transpositions, exchanges of onsets and codas, or even whole-syllable transpositions (though Vousden, Brown, & Harley, 2000 show that syllable-based speech errors are not just word onset errors, as suggested by Shattuck-Hufnagel, 1987). If speech errors are viewed as breakdowns in planning, differences in speech errors in different languages suggest that different units go awry during speech, and therefore that planning units differ between languages. For example, Mandarin Chinese is said to have syllable-level planning units due to syllable transpositions in speech errors (T.-M. Chen, Dell, & Chen, 2004; O'Seaghdha & Chen, 2009).

The other standard method of probing planning units is through priming experiments. In syllabic priming tasks, subjects must produce a target word that does or does not match a syllable prime. In an early study of this type, Cutler et al. (Cutler, Mehler, Norris, & Segui, 1983) asked French speaking subjects to listen for a CV or a CVC sequence in target words that began either with a CV or a CVC syllable. For example, subjects would be asked to listen for 'ba' in 'balade' and 'balcon' (among other words). These two target words differ in syllabification: ba.lade begins with a CV syllable, and bal.con begins with a CVC syllable. Cutler et al. found that subjects were faster at finding CV sequences in ba.lade-type words than in bal.con-type

words, and faster at finding CVC sequences in bal.con-type words than in ba.lade-type words. In other words, subjects were more primed to recognize whole syllables than strings of phonemes.

Parallel studies have now been performed in a variety of languages. In naming studies, where subjects pronounce a word displayed on a screen, English speakers are faster to begin production when primed with the word's first syllable than when primed with a partial syllable or a unit larger than the syllable.

However, there are two types of exceptions to these findings. One is that English ambisyllabic words like *balance* are named at the same speed following a CV or CVC prime. The authors claim that this is not a coincidence; the flexibility in priming unit shows that these words are best analyzed as having either an initial CV or CVC syllable (Ferrand, Segui, & Humphreys, 1997). These results have a parallel in Tilsen (2009), in which talkers asked to stop producing a sentence immediately upon seeing a signal stopped more slowly before stressed syllables than before unstressed syllables. There seems to be an interaction between stress and units of planning such that stressed units are longer or more cohesive. Another exception is that, in some naming tasks, English and other Germanic languages fail to show the syllabic effects first demonstrated in French. There are even studies in these languages providing evidence that speech production is organized on a sub-syllabic level. For example, picture naming is faster in English when the picture to be named is shown alongside a distractor that shares a syllable onset, for example, *can* and *cat* (Morsella & Miozzo, 2002). These two words share a word-initial [k], but are otherwise quite different phonetically. The vowel in *can* is highly nasalized while the vowel in *cat* is not, and the nasal coda /n/ is, featurally and acoustically, quite different from the coda /t/. For these words to affect each other, some vestige of their phonemic composition must be available when they are retrieved. Apparently, Germanic languages can make use of subsyllabic planning units in naming tasks.

Failure to find consistent crosslinguistic syllable priming effects have called the original Cutler et al. results into question, and indeed, follow-up studies in French have had mixed results. For example, Ferrand et al. (1996) replicate the Cutler et al. whole-syllable priming for naming of French two-syllable words, and Laganaro (2006) finds syllable frequency effects in picture naming latencies for Spanish and French, but Brand et al. (2003) replicates the procedure of Ferrand et al. and fails to find syllable effects, even in French. Further discussion of this line of research can be found in Schiller, Costa, & Colomé, 2003. In fact, only one language has been consistently shown to generate a syllable effect across experimental paradigms: Mandarin Chinese (T.-M. Chen et al., 2004). The authors speculate that this may be due to the relatively small number of different syllables in Mandarin relative to English. Mandarin has about 400-1200 syllables, depending on whether syllables with different tones are considered to be different syllables, whereas English has about 10,000.

Studies of Japanese, which is said to be planned in moraic units, serve as a useful counterpoint to the Chinese and Indo-European studies. These studies have found

several pieces of evidence supporting moraic planning units. First, Japanese speakers can begin saying words faster when they have been primed with whole moras than they can when primed by C or Cj subsyllabic units (Kureta, Fushimi, & Tatsumi, 2006). In addition, Japanese speakers are prone to speech errors that preserve moraic structure. For example, transposition errors tend to replace bimoraic syllables with bimoraic syllables, and non-syllabic morae with non-syllabic morae.

1. ko-ku-ro.o do.**o**-ro.o → ko-ku-ro.o do-**ku**-ro.o 'Kokuroo and Dooroo' (names of railway workers' labor unions)

2. zyuu.go pa**a**.sen.to → zyuu.go pa**n**.sen.to 'fifteen percent'

   (Kubozono, 1989)

Finally, Kubozono notes that most Japanese stutterers are said to stutter with moraic-sized units. That is, if the target were the word *soo.si.ki* 'funeral', a stutterer is more likely to say *so- so - soo.si.ki* than *soo- soo- soo.si.ki* or *s- s- soo.si.ki* (Kubozono, 1996, p. 80).

Most early interpretations of these studies suggest that differences in syllable chunking are driven by the structure of the language, and that the effect is language-specific. Under this interpretation, Mandarin Chinese is planned in syllable-sized units, while planning in Germanic languages and French might be planned in syllable or phoneme-sized units.

Additional crosslinguistic evidence against planning units of a fixed size comes from Pluymaekers, Ernestus, and Baayen (2005). This paper analyzes a corpus of conversational Dutch speech, looking specifically for factors influencing the duration of 8 very common adverbs as they are repeated over the course of a conversation. It is well known that repeated words get shorter as they become given informatoon in a conversation. But surprisingly, the authors of this study find that words are not reduced uniformly. Instead, the subword components that are reduced depends on the word. They argue that if reduction happens on the scale of whole planning units, then those planning units are not syllables, morphemes, or whole words. The most likely explanation is that, while a speaker's native language predisposes him or her to look for syllables or phonemes, the size of planning units is a function of speaker, not language.

A reasonable way of unifying these studies was proposed by Berg & Al-Jawad (1996), who argue that more than one unit size can be activated in a single language, and that either phonemes or syllables might be more important at any given moment in time. As evidence, they compare German and Arabic speech errors. Though Arabic contains both between-word and within-word errors, its errors are less likely to respect syllable boundaries or syllable position than errors in German. In their view, a typical German speech error would be *pessimischtis* [pɛsi:mɪʃtɪs] for *pessimistisch* [pɛsi:mɪstɪʃ], in that two coda consonants, [s] and [ʃ] are exchanged within a word.

A typical Arabic speech error would be /burgdaan/ for /burdgaan/, an exchange of two stops in adjacent syllables with different syllable positions; [d] is a syllable onset and [g] is a syllable coda. Similar to a major model of speech production proposed by Levelt et al., they claim that initially, words are retrieved as a string of phonemes. Arabic, which has non-concatenative morphology, must add an initial step in which words stored as a string of about 3 consonants are interspersed with vowels to create a word. For example, in this initial step, the root *k-t-b* might become the word *katab* 'he wrote', *kitaabun* 'book', or *maktab* 'office'. Errors at this early stage may not respect syllable boundaries since the speech production system has not yet assigned them. Later, words are syllabified and strung together with adjacent words. At the syllabification stage, between-word errors are more likely, and those errors will respect syllabic boundaries.

Summary: planning unit sizes

Psycholinguistic evidence is split on the mental units that compose speech production. Some studies find functional units on the level of the syllable, and some find sub-syllabic units. Experimental studies suggest that units may be of different sizes in different languages. For example, English speech errors show effects of both phonemes and syllables on word-naming (O'Seaghdha & Chen, 2009). Chinese speakers' speech errors, on the other hand, have pointed only to syllabic units (O'Seaghdha & Chen, 2009), and Japanese speakers' errors suggest planning in moraic units (Kureta et al., 2006).

Taken together, this evidence suggests that talkers are able to plan speech on multiple levels. This review, and this dissertation, focuses on vowel representations, and specifically, how they change on the basis of input from themselves and others.

Vowel targets

Because talkers (generally) produce vowels with vocal tracts, a good deal of early phonetic research was devoted to studying the physiology of the vocal tract and how its shape changes over time, with the hope that vowel targets would arise naturally from studying the apparatus producing them. Yet despite a century of trying, it is still uncertain what we are listening for when we listen to vowels.

There is some evidence that we listen to harmonics; to vowel dynamics in particular frequency regions; to formants; to broadband formants; to narrow-band formants.

Chiba and Kajiyama, in their seminal 1941 book meausred vocal tract area in men, women, and children, derived the area function of the vocal tract during vowel production. Their model forms the foundation of perturbation theory, in which changes to constriction location alter formant values, and continues to inspire models of vowel production through the present day. Their model was famously used by Fant in his Acoustic Theory of Speech Production, which proposed a mapping between an

area function of the vocal tract and its associated vowel formants calculated from developing theory about acoustics. Many modern synthesizers that must translate between constriction locations and acoustics (e.g., Story, 2005) have their roots in this theory.

For the purposes of analysis, it is common to think of a vowel target as a single configuration of the vocal tract, and two formants, F1 and F2 (and occasionally F3, especially for rhotic and rounded vowels). Under special conditions – male speakers of Midwestern American English producing *hVd* words in isolation – F1 and F2 are almost uniquely able to separate vowel categories (Peterson & Barney, 1952).

The success of vowel separation in early citation form production studies notwithstanding, vowel perception is a paradox. Talkers produce vowels with a vocal tract, but by measures directly derived from this vocal tract, vowels are neither uniform across speakers nor unique for each vowel. Formant frequencies, the resonant frequencies of the vocal tract, are characteristic of a particular person, word, and speaking style: change any one of these variables and formants change in seemingly unpredictable ways. Constriction locations determine formant frequencies and suffer from the same variability between men and women, consonant contexts, styles, etc. Indeed, there is substantial speaker and gender dependence in formant values, especially in F2 (see Nearey, 1989 for discussion). The paradox is that we understand each other to be producing the same vowels, when in fact there is nothing obviously the same about them in either the acoustic or the articulatory domain. This troubling fact has led to a concerted effort to find *invariants* in vowel acoustics and articulation.

Peterson and Barney's landmark (1952) study measured formant frequencies in *hVd* words produced by men, women, and children who spoke a Midwestern variety of American English. Formants produced by these groups have been used as a benchmark against which to compare American vowel formants ever since. Potter and Steinberg did the same for 25 speakers and, surprisingly, found that vowel formants in the three groups overlapped: when a convex hull was drawn surrounding all productions of all vowels, there was a small amount of overlap between some of the vowels, even in citation form *hVd* words. Some of this overlap can be resolved by distinguishing vowels not by the first two formants, but by ratios of those two formants. Potter and Steinberg suggest a function of frequency and amplitude, and demonstrate that fundamental frequency might play a small role in vowel identification. They are able to get separation of produced formants in citation form vowels from men, women, and children using ratios of formants in Mels (a transform to bring Hertz measurements closer to what the inner ear transmits). Syrdal and Gopal innovated a different normalization scheme, based on measurements in Bark units (a transform in which each unit excites an equally sized portion of the cochlear membrane), and separated them by Bark differences (F2-F1, F1-F0). Because Barks are roughly logarithmic relative to Hertz, this is similar to $(\frac{\mathrm{F2}}{\mathrm{F1}}, \frac{\mathrm{F1}}{\mathrm{F0}})$ in Hertz. Miller defined a sensory reference based on the geometric mean of F0, and created a normalized formant space using a

function of this sensory reference and log transforms of vowel formants.

Not all researchers have taken a formant ratio approach. Traunmüller (1988) claims that women's vowel spectra are not simply shifted and stretched versions of men's vowel spectra. Bladon et al. (1994) match entire Bark vs. sone spectra to vowel templates, and classify vowels based on the degree of match.

Formant ratios may be able to distinguish English citation form vowels across many speakers, but to deal with vowels in a variety of consonant contexts and in casual or rapid speech, they are not sufficient (Johnson, Flemming, & Wright, 1993; Johnson, Strand, & D'Imperio, 1999). That is, the vowel in *head* produced by women has the same formant frequencies as the vowel in *had* produced by men, yet talkers have no trouble identifying the words in these two overlapping groups. In some cases, vowels may be unambiguous because they have unique dynamics (Nearey, 1989; Strange, Jenkins, & Johnson, 1983).

By transforming formants into a more complex function of formants, it is possible to separate vowels based on perceptual and production criteria. Using vowel production from a number of speakers, Broad (1976) creates "equivalence regions" for vowel categories. Others have created categories based on psychophysical regions. The psychophysical method generally involves creating two synthetic vowels composed of some number of formants (between one (Traunmüller, 1981) and five (Pisoni, 1973)), and synthesizing a continuum that interpolates between formant values for the two endpoints. Subjects are asked to identify what vowel they hear at each step along the continuum.

There are two primary conclusions drawn from these studies. First, vowel perception is relatively categorical. There is a critical point along a synthetic continuum where perception flips such that most speakers stop identifying the vowel as one endpoint and start identifying it as the other endpoint (Cohen, Slis, & 'T Hart, 1963), and discrimination is better *between* vowel categories of American English than it is for distinctions of equal size *within* categories. Second, vowel boundaries from real speech are not linearly separable in any interpretable space. It is always possible to choose some set of dimensions wherein all vowels on one side of a line would be identified as one vowel and all vowels on the other side would be identified as another vowel, but the dimensions that one would be forced to use are not meaningful.

Although vowels may not be heard in terms of particular formants or formant ratios, there is compelling evidence that individual speakers strive to maintain particular formants in particular vowels, at least under laboratory conditions. Talkers wearing bite blocks preventing them from achieving their normal constriction pattern for /u/ will often change their articulation of that vowel so as to maintain standard /u/ formants (Perkell, Matthies, Svirsky, & Jordan, 1993). These "trading relations" enable many speakers to switch between two articulatory configurations that yield approximately the same acoustic signal based on "predictive simulations" (Lindblom, Lubker, & Gay, 1979) of how the intended articulation will sound (of which more later). This phenomenon has also been documented with American English /ɹ/ (Guenther et al.,

1999).

Similarly, Kelso and Tuller (1983) exposed subjects to "extremely deprived" sensory feedback by having them wear a bite block, then giving them an anesthetic to block transmission of jaw position and movement; spraying the inside of their mouths with Xylocaine, an oral anesthetic, reducing somatosensory feedback; and having them wear headphones playing loud white noise, masking their auditory feedback. Surprisingly, subjects produced the same vowel formants under these extreme conditions that they produced without sensory blockage. This evidence suggests that speakers have motor programs that are robust and flexible under a multitude of conditions. Neither auditory nor somatosensory feedback is crucial to accurate articulation.

Perception research has also attempted to isolate the necessary components of vowels, with the expectation that those components, at least, are essential to vowel identity. One such approach examines vowel production over time in adults with hearing loss. After adult-onset hearing loss, speakers quickly lose the ability to control the amplitude and spectral tilt of speech sounds, but the ability to produce consistent, correct vowels and consonants remains for many years (Perkell et al., 2000).

Another approach is to ask listeners to identify synthetic vowels that vary along a dimension thought to be important to vowel identity (Hose, Langner, & Scheich, 1983). The trouble is, there are no real invariance here, either. Listeners can hear vowels in stimuli that have been impoverished in surprising ways, including vowels with a single formant (Traunmüller, 1981), vowels without a fundamental frequency (Fahey & Diehl, 1996; Fahey, Diehl, & Traunmüller, 1996; Diehl, 2000), and vowels missing key portions of the middle of the vowel, so-called "silent center" vowels (Strange, 1989).

Furthermore, patterns in vowel boundary identification within a language seem not to hold across languages. Speakers of languages with different vowel inventories categorize vowel boundaries differently, as measured by the performance of American English speakers against speakers of Swedish (Stevens, 1969), Hindi (Hawkins & Stevens, 1985; Beddor & Strange, 1982), and French (Gottfried, 1984).

Other evidence for the composition of vowel targets comes from asking speakers to exaggerate vowel sounds and examining what characteristics of the vowel they exaggerate. So-called *clear speech* is associated with a large (17%) improvement in intelligibility over conversational speech (Picheny et al., 1985, 1986). It tends to be implemented by producing vowels with extreme formants, increasing the formant distance between adjacent vowels. Some familiarity with the phonology of the native language is necessary for gaining information from these formant exaggerations; non-native speakers of English generally do not experience the same improvement in intelligibility due to clear speech (Bradlow, 2002). Similarly, subjects asked to synthesize standard American English vowels prefer to use extreme formant values (Johnson et al., 1993). These lines of work suggest that speakers understand what the relationship between formants in their native language ought to be. Speakers prefer

to increase intelligibility by increasing the distance between adjacent vowels, even when the resulting formants are no longer near the formants that are most typical of that vowel.

Given that vowels fail to correspond to formant or formant ratio invariants, it is worthwhile to wonder whether vowels are associated with acoustic invariants at all. Gestural theories of speech perception and production take an opposing view, that speech is heard in terms of movements of the articulators and configurations of the vocal tract. Gestural theories of speech production propose that there is a mapping between *tract variables* specifying the locations of constrictions in the vocal tract and the articulators that need to move in order to achieve those configurations. The set of required articulatory movements, called gestures, are planned in sequence. Indeed, in some situations, the gestures required to produce speech sounds are more consistent across speakers than the acoustics of those speech sounds, which depend on the length and physiology of the vocal tract. The timing of the elements of a sequence of gestures leads naturally to boundary effects between phonemes (Fowler & Saltzman, 1993). Browman and Goldstein explain that their Articulatory Phonology framework, which views speech as sequences of gestures, is useful for accounting for a number of phonological alternations such as the unaspirated /p/ in /sp/ clusters, as well as coarticulatory effects such as /k/ fronting in front vowel contexts (Browman & Goldstein, 1992).

Thus, though it is common to simplify our analysis of English vowels by viewing them as monophthongs, the success of gestural formulations of vowels demonstrates that vowels in a word context are never steady state. It is more accurate to think of formants as "temporal multidimensional ribbons" (Perrier, Lœvenbruck, & Payan, 1996, p. 57), waving through articulatory space and time.

Vowel dynamics

Articulatory Phonology has inspired many researchers to view vowel targets as spectrotemporal rather than spectral. There are a number of reasons for this. One is that vowels are not produced in isolation; they are produced as part of larger syllables, words, and sentences. Articulating all of these sounds in tandem with a single set of articulators causes coarticulation; every sound is influenced by surrounding sounds. Even if vowels have a single set of formant targets, a vowel is really the process of ramping up to this target and then the process of moving from the target to the next sound. It is also possible that vowels are not actually stored as sets of single formant targets, and that this was merely a simplifying assumption of early vowel research. As phonetic research has recorded more precise variation, evidence has mounted for variation in even productions of apparent monophthongs.

Vowel dynamics have been documented in a number of dialects of American English and crosslinguistically as well. For example, Hillenbrand, Getty, Clark, and Wheeler (1995) showed that all vowels except /i/ in Midwestern American English

show movement over time. For example, /ɛ/ and /æ/ are quite close to each other in static formant space, but their offglides differ: the difference between F1 and F0 in /ɛ/ tends to decrease over time, whereas the difference between F1 and F0 in /æ/ tends to increase over time. Hillenbrand et al. also show that vowel recognition is aided in the context of changing vowel formants, and limited in the context of steady state formants. Based on this evidence, they claim that formant dynamics help to contrast similar pairs of vowels.

Outside of American English, Carré, Ainsworth, Jospa, Maeda, and Pasdeloup (2001) show that the speed of formant changes in diphthongs influences the perception of French listeners. Not only does the shape of the vowel trajectory affect vowel goodness judgments, but listeners actually hear different vowel sequences depending on the rate of formant change. Fast changes from /a/ to /i/ are heard as /ai/, while slow changes are heard as /aɛi/. Furthermore, Strange et al. (1983) show that vowels are always better identified in consonant context than in isolation, even when the consonant context forces vowel formants to edge away from vowel prototypes.

Aside from the dynamics intrinsic to vowels, additional dynamics complicate the formant trajectory. Because vowels are flanked by consonants and other vowels, *coarticulation* obscures their endpoints. Context affects acoustics in semi-predictable ways, but often serves to distort the vowel such that formant "targets" are never actually reached. This is especially common in fast speech.

Coarticulation

Another prominent feature of fast speech is coarticulation, the influence of one segment on surrounding segments. To understand why coarticulation happens, consider the production of a CV sequence. As shown in Figure 2.1, CV production requires releasing a stop closure, then articulating the following vowel. The shift from C to V requires tongue and jaw movement and is not instantaneous; as a result, a CV syllable can be split into a C release, a transition from C to V, and a V. Research on the formant values and articulation of the C burst and the V indicate that the two are not independent; midpoint formant values of a vowel depend on the surrounding consonants, and the formant values following the stop release bursts depend on the surrounding vowels. However, vowels don't change as much due to consonant environment as consonants change due to vowel environment (Keating, Lindblom, Lubker, & Kreiman, 1994). The co-influence of vowels and consonants is taken as evidence for suprasegmental units of planning, and has spawned several theories about how these coarticulatory effects arise.

There are at least two ways to conceive of coarticulatory planning. One is that segments are planned one at a time. The simplest method of planning one segment at a time is to posit that each phoneme is stored as a group of context-sensitive allophones (Wickelgren, 1969). This method requires a large amount of storage space and does not permit abstraction of a phoneme over multiple contexts. A more prac-

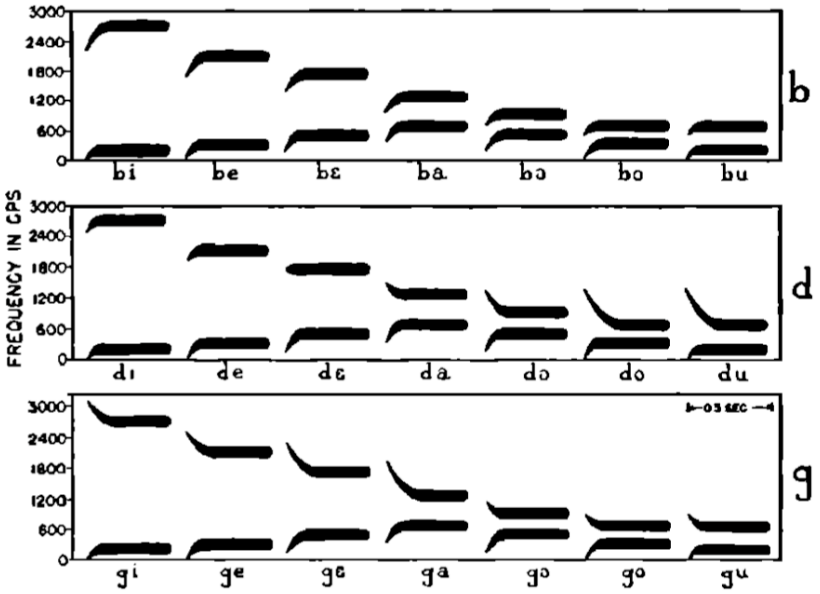


Figure 2.1: Vowel formant transitions in American English. Reprinted with permission from Delattre, Liberman, & Cooper, 1955. Copyright 1955, Acoustical Society of America.

tical method is to compute coarticulation online, as in Henke's "look-ahead" model (Henke, 1966). In this model, words are strings of phonemes, and phonemes are bundles of features. During coarticulation, phoneme features can spread backward from upcoming phonemes back to the current phoneme. Some subsequent studies provided evidence for this model (Farnetani, 1999; Sussman & Westbury, 1981). Another view traces coarticulation not to prediction of upcoming sounds, but rather to a compulsion to save effort in production (Lindblom, Sussman, Modarresi, & Burlingame, 2002). In Lindblom et al.'s view, coarticulation is a type of reduction, expected to be greater when the listener can handle it (i.e., when the words aren't confusable). Both of these models depend on a segmental representation, but would also require some sort of suprasegmental representation because coarticulation is not *a priori* predictable. Every language phonologizes coarticulation such that some phonemes are articulated with greater coarticulation than necessary. Thus coarticulation is not simply a natural consequence of a sequence of gestures: it is affected by neighborhood density and frequency and remains even during clear speech (Farnetani, 1999; Bradlow, 2002). It is also language-specific. To take one example of many, nasalization in Catalan does not spread to the vowel preceding nasal consonants (Shockey, 2003, p. 19)

It is also worth exploring what coarticulation would look like if speech were planned on the syllabic level. Coarticulation in this case would have to be built into the syllable rather than computed on the fly, though online computation would still be

required at syllable boundaries. The DIVA model of speech motor control currently subscribes to this view. It assumes that speech targets encompass entire syllables, and that infants learning to speak are learning to hear and say whole syllables. Conceptually, Öhman (1967) had a similarly suprasegmental view of coarticulation. He proposed that production is planned on the suprasyllabic level: in the case of /idi/, for example, there is coarticulation between the /d/ and the /i/ caused by both movement from V1 to V2, and movement of the tongue tip for articulation of the consonant. He made the observation that in VCV sequences, the vowels flanking the consonant affect each other even though the intervening consonant might have wiped away all previous vocalic influence. This is a suprasegmental, but hierarchical model of coarticulation.

Another possibility is a gestural theory of speech production, in which the relative timing of successive gestures gives rise to coarticulation. In Saltzman and Munhall's (1989) model, planning is phrased in abstract constriction targets, which are implemented in terms of gestures. The dynamics required to plan production this way involve activation of multiple gestures at the same or nearby points in time, and overlap of activations might cause the point of constriction to be a blend of two adjacent constrictions, or might cause target over- or undershoot. For example, if a vowel gesture is cut short by starting the following consonant gesture, a speaker would not quite reach the vowel's prototypical formants. Because there is not a one-to-one relationship between gestures and phonemes, gestures activate and deactivate over the course of multiple phonemes, and coarticulation is highly context-dependent.

Coarticulation affects the position of the tongue body during consonants as well. Experiments using electromagnetic midsagittal articulometry (EMMA), in which alternating voltages induced in small metal spheres appended to the surface tongue are used to track their position over time, find that the tongue body is further front for consonants in /i/ context than those in /ʌ/ context (Fowler, 1994; Fowler & Brancazio, 2000). Furthermore, coarticulation is essential to comprehension: phoneme strings synthesized without coarticulation cannot be understood (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

All in all, vowels in running speech are tremendously rich, high-dimensional signals that hold only a family resemblance to each other. Phoneticians have done an admirable job of starting to sort them out. Yet there is still considerable debate on the objects of perception and the necessary components of targets in vowel production.

This dissertation investigates the components of vowel targets by exploring how they are learned and changed. This process logically entails taking in an external signal, comparing it to an expected signal (generated internally), and making adjustments to account for discrepancies. In the case of other voices, the comparison is mediated by a number of social variables, but in the case of one's own speech, no theory of mind or sociology is necessary to explain target change. The system that maps one's own input to output is called the *speech motor control system*.

Target change

However articulatory targets are specified, they are subject to change over time for a variety of reasons. One of the most salient is development. The set of muscle movements that generate an /a/ or a /k/ in a 6-month-old will not produce those sounds in a 6-year-old. This is because both the absolute and the relative sizes of the articulators change substantially over the course of development (Vorperian et al., 2005; Vorperian & Kent, 2007). For instance, at six months, the posterior portion of an infant's vocal tract is at 40% of adult size, but its anterior cavity is 75% of adult size! As these cavities change in relative proportions, constrictions at the same horizontal locations in the vocal tract will correspond to wildly different formants. Furthermore, because the vocal tract lengthens with age, changing the harmonic structure of vowels, a 6-year-old cannot physically produce vowels with the same formants as those of a 6-month-old. Both acoustic and somatosensory components of vowel targets *must* change during development.

There is no consensus on how vowel target change happens, but the process almost certainly involves self-monitoring. To see why, imagine a small child with an acoustic and a somatosensory target for /u/. One day, the child executes the articulatory plan that previously produced /u/ formants and ends up producing a vowel with unexpected formants, perhaps an unusually low F2 due to the child's growing pharynx. To keep from being misunderstood or mocked, the child needs to adjust that vowel's articulatory plan fairly quickly and make it align with the expected formants. But at some point, as the child's vocal structures change, it will become impossible to produce exactly the same set of formants that were produced for /u/ before, and the acoustic target will have to change as well. In reality, the two probably change in tandem, based on some balance of intelligibility and effort (Lindblom, 1990). Adjustment requires access to (1) the formants and articulatory movements that were just produced, and (2) an *expectation* for those formants and articulatory movements. The *speech motor control* system tracks and compares these two pieces of information.

There are two competing theories for how a talker's expectation for what he is about to hear is generated. One, exemplified by the DIVA model of speech production (Guenther, 1995; Guenther, Hampson, & Johnson, 1998; Guenther, 2003; Bohland, Bullock, & Guenther, 2010), relies on a feedforward model learned during early infancy. This theory supposes that as infants babble, they store the relationship between motor commands and the acoustic signal that those commands generate, and subsequently use this mapping to convert between articulation and acoustics (and vice versa) throughout life. Once syllables in the native language are learned, those syllables are matched to a set of possible articulations and acoustics, permitting a straightforward lookup of expected acoustics from an intended syllable.

The opposing theory supposes that talkers store an *efference copy* of the signal sent to their articulators (Houde & Nagarajan, under review). This is a copy of the

motor commands that were sent, along with a "corollary discharge" indicating that the movements were self-initiated (Sperry, 1950; Blakemore, Goodbody, & Wolpert, 1998). This theory proposes that the efference copy generates an expectation for the phonemes or syllables that were just produced by sending motor commands through an internal model of the vocal tract. An internal model based on efference copy creates a much more specific expectation than does a feedforward model because the feedforward model is generated from an intended syllable, and does not differentiate between different productions of the same syllable, whereas a model based on efference copy generates an acoustic expectation, blind to the intended syllable, based only on motor commands.

Thus far, no studies have tracked expectations and articulations in small children as they develop. However, a number of studies with adults have perturbed feedback from the somatosensory, auditory, and visual systems. Two separate phenomena have been documented when subjects experience manipulated feedback. The first is a behavioral response during manipulated trials, called compensation. Subjects generally compensate by opposing the feedback manipulation, though feedback following has also been documented, particularly in pitch perturbation experiments. The second phenomenon is adaptation, "a semipermanent change or perception or perceptual-motor coordination" (Welch, 1986, p. 24-3). Some consider adaptation to be a a demonstration of learning.

In fact, subjects eventually adapt to even the grossest changes to the visual field, for example, wearing glasses that make the world appear upside-down, well enough to ski and ride a bicycle in traffic (Welch, 1986, p. 24-16). Occasionally, adaptation happens piecemeal; one subject wearing glasses that flipped the visual field horizontally reported seeing a car on the correct side of the road, but seeing the figures on the license plate in reverse.

In response to less disruptive changes to vision, compensation for the visual shift is not complete. Subjects who wore prism glasses distorting their visual field and reached toward a target (they were unable to see their arms or hands while reaching) compensated a maximum of 85% for the visual shift (Welch, 1971).

Visual targets clearly have visual components, but the components of speech targets have been more difficult to uncover. Studies of speech targets using perturbation paradigms examine whether talkers respond to disruption or distortion of a source of feedback during speech. Such experiments have led to a consensus that vowel (and fricative) targets contain expectations in the somatosensory and auditory domains. Some of the initial studies in this field altered auditory feedback by delaying it (e.g., Fairbanks, 1955). These studies found that speech was highly disrupted by the delay, suggesting that auditory feedback is crucial to smooth and accurate execution of articulatory plans. Other early studies on brief perturbations to somatosensory feedback revealed that talkers have expectations for the relative positions of some articulators, and for the absolute positions of others. When a paddle is briefly applied to a speaker's upper lip during the production of a sequence like [aba], the

most straightforward response would be to increase the force pulling the lower lip up, opposing the force from the lip paddle. Speakers do not do this. Instead, they lower their upper lip, maintaining the bilabial closure (Abbs & Gracco, 1984; Gracco & Löfqvist, 1994; Munhall, Löfqvist, & Kelso, 1994). There are two ways to account for compensation by the upper lip. One explanation is that subjects are trying to maintain the acoustics of the intended syllable, and the best way to make the /b/ in /aba/ sound like a stop is to lower the upper lip. Another explanation is that subjects have *somatosensory* expectations for intended syllables; that is, they know what those syllables are supposed to *feel* like. In the case of /aba/, they know that their jaws should feel open during the vowels and that their lips should come in contact during the /b/. Maintaining somatosensory information about this syllable would likewise require compensation for the lip paddle in /aba/.

Further evidence for auditory expectations comes from subjects whose palate or dentition is temporarily deformed, perturbing somatosensory feedback. Honda & Fujino, 2002 asked two subjects to wear an inflatable palate and produce /ʃa/ or /tʃa/, either with or without auditory masking[1]. The palate was inflated during randomly selected trials. Later, speakers were recorded after practicing speaking with the inflated palate, and immediately after the palate was suddenly deflated. A group of listeners identified productions of /ʃa/ or /tʃa/ produced by speakers either after practice wearing the inflatable palate, or just after the palate was suddenly deflated. Listeners generally misperceived the first production after the palate was deflated, but correctly identified all subsequent productions. The fact that the reaction times were very short, and that substantial compensation occurred with masked feedback, suggests that somatosensory feedback initiates initial adjustments to articulation in the presence of the inflatable palate. That intelligibility improves after the first token suggests that auditory feedback is also used to adjust articulatory plans. That subjects are able to adapt and produce intelligible /s/'s and /ʃa/'s under these conditions suggests that talkers have expectations for what syllables should sound like, and have a good enough map of articulation to acoustics to accommodate gross changes to the shape of the vocal tract.

Similarly, in McFarland and Baum (1995), subjects produced vowels while wearing a bite block which prevented them from producing vowels and consonants with their typical articulations. The authors looked for articulatory compensation by measuring the formants of vowels that subjects produced and the spectra of fricatives and stops that they produced in order to determine whether subjects were compensating for the bite block by maintaining acoustics and changing articulation. Subjects did compensate, though not immediately. Compensation for vowels grew better over several minutes, and compensation for nonvowels continued to improve over the course of 15 minutes. This is evidence for acoustic targets that are under tight control,

[1]During auditory feedback masking, subjects wear earphones that play a loud hissing noise as they speak, preventing them from hearing their own voices.

especially for consonants.

Subjects also make heavy use of auditory feedback when their dentition is altered (Jones & Munhall, 2003). In this experiment, subjects' upper incisors were prolonged with a dental prosthesis that increased tooth length, and speakers' /s/-productions were recorded with auditory feedback masking and without. Initially, the longer incisors lengthened the front cavity, lowering the center of gravity of the /s/ productions. Auditory feedback and practice increased the quality of /s/ productions, as judged by naive listeners.

Other studies have shown that somatosensory feedback is critical to speech targets as well. In a representative experiment, subjects were strapped into an apparatus holding their heads in place. They were asked to say a single-syllable word, either 'row' or 'straw'. During some trials, a load was applied to the jaw, pulling it to the left or right. The pull was unnatural but did not change the acoustics of the word being produced. Most subjects pulled their jaw in a direction away from the load (for example, pulling right when the load pulled left), about halfway toward neutral position (Nasir & Ostry, 2006). Most subjects also compensate for jaw loading during normally-voiced speech or silently-mouthed speech, but not during nonspeech (Tremblay, Shiller, & Ostry, 2003)[2]. Compensation for jaw perturbation is strong evidence for somatosensory expectations in speech movements.

Somatosensory expectations are common in motor control in other domains as well. Related perturbation experiments have investigated the effect of altered feedback on keystrokes in practiced typists, demonstrating a clear mismatch between automatic error detection and ability to recognize errors. In these experiments (Gordon & Soechting, 1995; Logan & Crump, 2010), typists see a word appear on a screen and are prompted to type it. Usually their keystrokes appear veridically, but occasionally the computer inserts an error when the typist had typed correctly, or corrects an error that a subject made. Because a typist typically slows down after making an error, this study was able to measure error awareness by checking whether typists slowed their typing after an error they did not commit (an inserted error), or failed to slow down after an error that was corrected. The study found that typists slowed down after errors they made, even if they were corrected by the computer, and did not slow down after inserted errors or correct keystrokes. However, typists' explicit awareness of their errors did not match their patterns in typing speed. Although typists slowed down after making errors that were corrected, about 40% of typists reported that they had not made an error. The authors conclude that typists must have an internal mechanism for monitoring their intended keystrokes that does not depend on visual feedback from the screen.

In the arm motor control literature, somatosensory information is often called proprioceptive information because it comes from internal estimates of joint angles, velocities, and positions. In two experiments, Sober and Sabes (2003, 2005) inves-

[2]Oddly, one-third of subjects did not respond at all.

tigate the integration of visual perception and proprioception in arm movements. Subjects in their experiments lay their arm on a low-friction sled on a table. Because the table and their arm is blocked by a drape, all of their visual feedback comes from a projection screen. Rather than showing the position of their arm on the table, the projection screen shows a white dot representing the location of their arm, and a red dot representing the location of a target. They are prompted to reach for a target, and after their arm has moved 5mm, the white dot disappears. During some trials, the dot representing their arm was moved 6cm to the left or right. Subjects made errors in their initial reach direction and velocity when this visual feedback was altered, and the pattern of errors suggested that visual and proprioceptive feedback were responsible for different parts of reach planning and execution.

Auditory feedback has been more difficult to perturb until recently. A growing body of work has altered feedback from vowels and select fricatives, showing that subjects compensate for altered auditory feedback (Houde & Jordan, 2002). There are two main paradigms for demonstrating compensation and adaptation in speech motor control.

In the first, formants are shifted uniformly over the entire stimulus word. With sufficient training, subjects compensate by opposing the shift in feedback. The first experiments using this paradigm required subjects to whisper, so that confounding feedback from bone conduction was not present. Houde and Jordan used this design to test adaptation to changes in auditory feedback. Subjects whispered and heard their whispered feedback played back to them in real time. Whispered feedback was altered using a vocoder such that [ɛ] feedback slowly changed to [ɑ] or [i]. Subjects compensated and adapted to the change in feedback to different degrees.

Subsequent experiments extended these results to voiced speech. For example, subjects in Purcell & Munhall, 2006a produced *hVd* words with veridical or altered auditory feedback. In their first experiment, altered feedback was ramped up to a maximum "hold", which was reached after 50 trials and held for an additional 15. On average, subjects compensated 29% for the altered feedback. The authors found the change point in subjects' pooled formant trajectories to decide when compensation began, and concluded that talkers began compensating at a feedback shift of 76 Hz. Experiment 2 used the same stimuli, but the number of trials at the maximum hold varied from 0-45 and was followed by 115 trials with normal auditory feedback. In this experiment, adaptation persisted into normal trials.

However, compensation seems to depend on the vowel being altered. Although the authors only measure compensation in F1, Purcell & Munhall, 2008 find differences in compensation for a ±200 F1 shift in auditory feedback for the vowels /i/, /ɛ/, /a/, and /u/. Compensation was greater for /ɛ/ than for /i/, /a/, or /u/. The authors suggest that mechanosensory feedback from the tongue is responsible for the smaller compensation for /i/.

The other paradigm for measuring compensation and adaptation is a brief *perturbation* occurring during the stimulus. In these experiments, the stimulus is usually

a sustained vowel rather than an entire word. Subjects in these experiments tend to start opposing shifts fairly quickly, within 100-225ms for F0, irrespective of the latency and magnitude of the perturbation (Burnett, Freedland, Larson, & Hain, 1998), and within 600ms for F1 (Purcell & Munhall, 2006b).

As in jaw perturbation experiments, there are substantial individual differences in compensation for altered auditory feedback. Some subjects compensate by changing their vowel production as much as is required to oppose the alteration, but others do not change their production at all. This individual variation is consistent with vowel representations that differ from speaker to speaker. But just as subjects in reaching experiments fail to compensate completely for a shift in their visual fields, subjects in auditory feedback shift experiments tend not to oppose the shift completely; a 50Hz change in production for a 200Hz shift in feedback is typical. Purcell and Munhall suggest that incomplete compensation may result from the relevance of the manipulated formants to vowel production; the more a speaker is accustomed to attending to an aspect of speech, the more they compensate for it. For example, Mandarin speakers compensated more for dynamic than for static tones (Jones & Munhall, 2002), and nontonal trained singers compensate less (and more slowly) for shifts in F0 feedback when they are producing a glissando rather than a steady state pitch (Burnett & Larson, 2002).

One way to use these experiments to explore the units of motor planning is to measure generalization of compensation to new contexts. If subjects plan individual phonemes, and one phoneme adapts to have new formants, that phoneme should be produced differently regardless of consonant environment. One study finds generalization of adaptation to other environments (Houde, 1997). If what is learned is instead a new internal model of the speech motor control system, one would expect that the entire acoustic-articulatory map would be affected and that generalization would spread to other vowels as well. Results are mixed on this point. In an experiment by Pile, Dajani, Purcell, and Munhall (2007), subjects produced 'hid', 'head', and 'hayed' with veridical feedback. Afterwards, F1 and F2 were shifted gradually until, when subjects produced the word 'head', they heard themselves saying /hæd/. After 40 trials with this maximum shift, they again produced 'hid' or 'hayed'. Although subjects compensated for the feedback shift in /ɛ/, formants in the surrounding vowels /ɪ/ and /e/ were unchanged. However, Houde (1997) finds that production differences in response to hearing a shift of [ɛ] in 'pep', 'peb', 'bep', and 'beb' generalized to 'gep', 'peg', and 'teg'.

One study has explored the effect of feedback shifts in units larger than words. S. H. Chen, Liu, Xu, & Larson, 2007 compared the response to an altered F0 during a vowel task, during which subjects produced the vowel /u/, and a sentence task, where subjects repeated a sentence that they heard (e.g. *you know Nina?*) This study found that subjects oppose the F0 shift more, and more quickly, in the sentence condition than in the vowel condition. For reasons that are not clear, subjects also responded more to downward than to upward pitch shifts.

The two findings that still need explanation – partial compensation for altered auditory feedback and individual differences in compensation responses – have led researchers to test a variety of hypotheses regarding influences on compensation for altered auditory feedback. Perkell et al., 2004 find that subjects with better hearing acuity compensated more for altered auditory feedback. Villacorta, Perkell, and Guenther (2007) find that subjects compensated more when they could discriminate F1 better or had low-variance vowel regions. Their experiment also demonstrates that subjects continued to compensate even when they could no longer hear themselves. In addition, Shiller, Sato, Gracco, and Baum (2009) show that there is a change in perception that results from adaptation experiments. Larson, Altman, Liu, and Hain find that altering both auditory and somatosensory feedback leads to greater compensation than altering auditory feedback alone. In their 2008 study, subjects produced a prolonged vowel before and after their vocal folds were anesthetized with a spray. When their F0 was perturbed, they compensated more when they could not feel their vocal folds, suggesting that veridical somatosensory feedback attenuates compensation for altered auditory feedback.

Summary: Target perturbations

Talkers normally hear and feel sounds within their expected range, and do not need to adjust their internal models or their targets. Early experiments altering vision demonstrated that talkers *compensate* for perturbations to sensory feedback. Bite block experiments showed that subjects compensate for perturbations during speech as well. Real-time feedback perturbations have fleshed out some of the details of these compensatory responses. When auditory feedback is altered, but somatosensory feedback is not, subjects adjust their motor commands to oppose the altered feedback and approach their acoustic targets. Changes in vowel production during language acquisition demonstrate that the target itself may change if new motor commands do not result in a sound within the old acoustic target region. When somatosensory feedback is altered, but auditory feedback is not, talkers adjust their motor commands to approach their expected vocal tract configuration. When auditory feedback is altered along with somatosensory feedback, subjects compensate more than they do when they have veridical somatosensory feedback.

Though a variety of factors related to hearing acuity are known to correlate with degree of compensation, it is still not clear why subjects compensate only partially, and why individual talkers respond so differently to altered auditory feedback.

Modeling changeable targets

There are multiple reasons that vowel targets might change over time. In accommodation, they change because new interlocutors provide new linguistic input and make old input more salient. During development, they change because the shape of

the mouth changes, and it's difficult or impossible to produce sounds within the old target range (Vorperian et al., 2005; Vorperian & Kent, 2007). Talkers can adjust either acoustic or somatosensory targets in these situations. A talker adjusting an acoustic target will try out a vowel with a new set of formants and decide whether the new sound is acceptable. A talker adjusting a somatosensory target will produce a new vocal tract configuration and determine whether the formants of the vowel produced with the new configuration are acceptable. As mentioned earlier, both types of targets are likely to be adjusted, with the proportion of adjustment in each domain chosen based on an optimization of understandability and effort.

A set of models that have been developed in the neuroscience and engineering literature can shed light on how targets adjust. Models of speech motor control have the goal of explaining how speech is planned and executed, just as speech processing models do, but they focus on the implementation and control of speech motor plans.

Just as in models of speech processing, there are disagreements on details, leading to several models of speech motor control, but all of them share the same skeleton. Input to these models comes in the form of subwords, usually phoneme strings or syllables. These subwords are associated with a learned, high-level articulatory plan. This plan is translated into motor commands, which take into account the current state of the articulators. Two types of output are generated from these motor commands: one, an expectation for what the executed subword ought to sound like, and two, audible speech from executing the motor commands. Feedback from the ears, with information about the speech sounds, and from mechanoreceptors, with information about the position of the articulators, can be compared with the computed expectation for both sources of speech information. Any mismatch between observed and expected feedback is fed back to the articulatory planning level and used to adjust subsequent speech.

Models of speech motor control

Most of the work explaining how compensation for altered feedback operates relies on control theory. *Control* is learning the computational inputs required for a system to generate the right outputs. This is a key problem in generating commands to execute any directed movement. Wolpert & Kawato (1998) use the example of drinking from a can to demonstrate that control of volitional movements is difficult. The series of muscle commands required to lift a can to one's lips varies depending on (1) internal variables such as the arm's current joint angles and velocities (also called the *state* of the arm), (2) external variables such as the moment of inertia of the can, the body's orientation relative to gravity, and torso acceleration (also called movement *context*), along with constants such as masses, moments of inertia, and centers of gravity.

There are two primary strategies for learning to perform complex tasks like these. One is to use a single controller that takes in all internal and contextual information

and generates the right commands. Computation is difficult with a single controller because there is a large number of possible scenarios and each one has to be hard-coded. This strategy is also error-prone because if the context changes before the system can adapt and change the response, the wrong output is generated.

Alternatively, one could choose a modular approach with multiple controllers, in which each module deals with a small set of contexts. Movement execution happens by estimating the current context, and activating controllers in proportion to the likelihood that the current context falls in their domain. The difficulty in this approach is learning which contexts should be assigned to which modules.

If applied to speech motor control, the Wolpert and Kawato model would posit that, prior to execution, the speech plan is divided into smaller components. Each of these small components would be assigned to multiple modules, and each module would take partial responsibility for generating one component of the speech plan. This method of dividing responsibility for a motor plan among multiple controllers is termed a "mixture of experts" model. Based on production error, the model learns to reassign responsibility for particular speech plans to different modules or to create new modules as necessary. This method generates an efficient coding of complex data. The modules that are eventually learned correspond to "primitives", basic motor units that are maximally useful for executing speech plans.

Because there is no way to know *a priori* what the speech motor "primitives" should be, it is critical to have an accurate way of measuring the error generated by using any particular combination of modules. Thus, one major challenge in implementing a model like this lies in deciding how the error is computed. Most models designed over the last decade utilize a so-called "homogeneous cost", in which all dimensions of the signal are equally important to generating the error. This would be useful if, for example, the model tracked F1 and F2 of vowels, and there was reason to believe that F1 and F2 error were equally important. In models using more complex signals, such as the auditory signal received by the speech motor control system, there are many potential dimensions to track, and some are more important than others. A model recently proposed by Liu and Todorov (2007) solves this problem by proposing a model where the *importance* of error dimensions are learned along with the weighting of the modules. Dimensions that vary a lot are considered to be more important and dimensions that are relatively constant are considered to be less important.

The Liu & Todorov model is consistent with previous research noting that corrections for moving a reach target late in the course of movement are not complete. The authors repeated this experiment, asking subjects to reach in the dark toward a target that might move at 100, 200, or 300 ms into the reach movement. They found that reachers undershot the target, even in trials where there was enough time to correct completely. Undershoot was greater in experiments where subjects were permitted to hit the target with whatever force they liked rather than touch the target gently. This is only possible in a model where subjects track high-level movement

goals, not trajectories or small subgoals, and treat stability as one of the costs to be minimized.

There are two speech motor control models that are based on these principles. One is the Directions into Velocities of Articulators model, or DIVA. In this model, units are syllables phrased in auditory, articulatory, and somatosensory coordinates (Guenther, 1995). The other is a State Feedback Control model (SFC), the model of speech motor control mentioned earlier which assumes that talkers store an efference copy of their productions. In DIVA and SFC, the incoming auditory signal is a crucial factor in ensuring that speech remains in the target region while speaking. Figure 2.2 shows the current structure of the DIVA model from Tourville & Guenther, 2010.

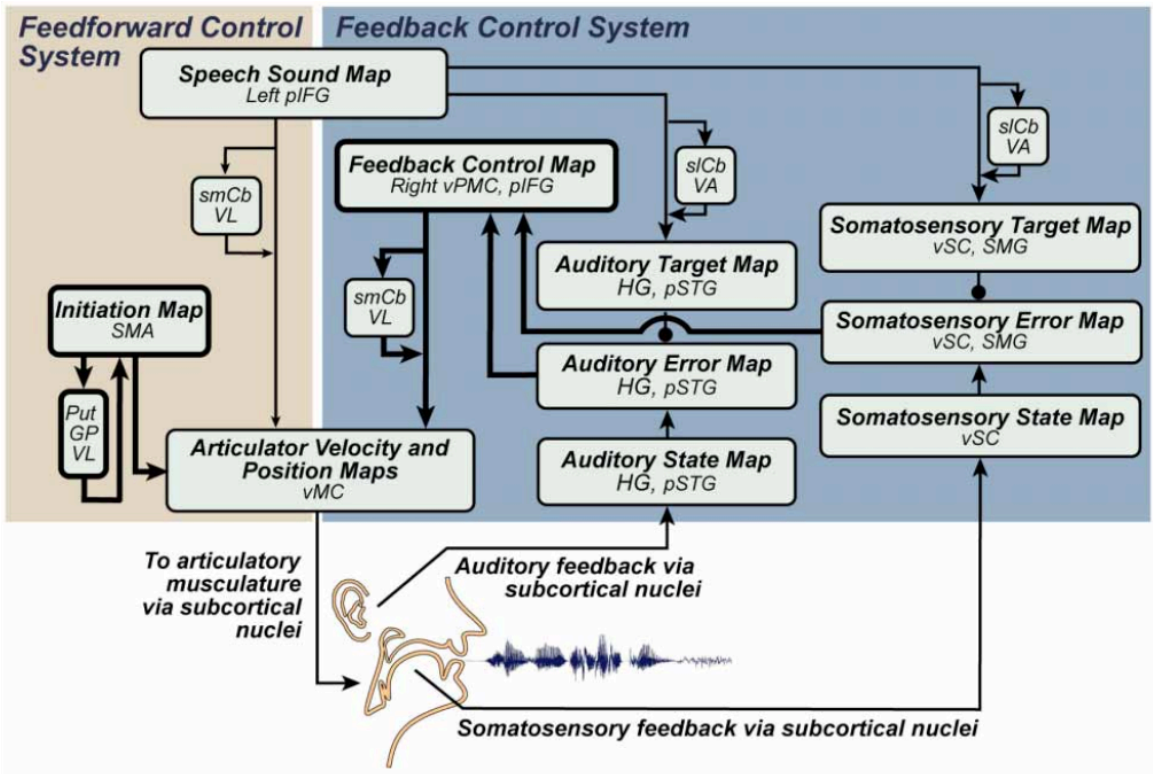


Figure 2.2: Current schematic of Directions into Velocities of Articulators (DIVA) model. From Tourville & Guenther, 2010.

In DIVA, target regions are formed by monitoring one's babbling during infancy and building a mapping between acoustics and articulation. As infants converge on motor programs that achieve consistent acoustic targets, they begin to associate the sound with the articulatory configurations and gestural plans that got them there. For this reason, acoustic and constrictional data compose two of the "reference frames" that Guenther posits that we monitor during speech. But these two reference frames constitute only a portion of the reference frames that are built into the model. The muscle length reference frame tracks the contractile states of muscles and their lengths and angles. The articulator reference frame is somewhat more abstract than the muscle length reference frame: it tracks the positions and trajectories of the individual articulators. The tactile reference frame tracks somatosensory feedback from mechanoreceptors in the vocal tract. The auditory-perceptual reference frame tracks acoustic signals that have been filtered by the ears and auditory cortex. That is, information about the locations of vocal tract constrictions is certainly present, but is no more important than any of the other reference frames. Importantly, in this model, the acoustic signal is primary, an idea that goes back to the Acoustic Theory of speech production (Fant, 1960).

Instructions for speech production in this model come from observing the current acoustic trajectory and a mapping, learned during infancy, between acoustics and articulation[3]. This central process gives it the name *directions into velocities of articulators*, or DIVA. As the articulatory plan proceeds, DIVA ensures that production is progressing as intended by comparing incoming information from all reference frames to target regions encompassing the range of acceptable productions in each of the reference frames. For example, if a speaker intends to say 'can', the DIVA model checks incoming auditory and somatosensory information from the ears and the muscles against a known range of acceptable sounds and articulatory configurations that correspond to 'can' and plans an adjustment if feedback falls outside of this so-called *target region*. In the DIVA model, only one copy of this mapping is used in the production of all syllables. Because there is only one mapping, the DIVA model predicts that changes to this mapping should generalize to all possible contexts.

A competing model, the State Feedback Control (SFC) model, is based on principles of optimal control. In SFC, the intended message is broken down into a sequence of *control laws*, planning units that are currently unspecified in size and composition. Because control laws are learned so as to optimally cover all known articulatory contexts, they may differ slightly in composition from person to person, and even within the same person as the L1 is learned. These models are commonly used in the arm motor control literature (Liu & Todorov, 2007). The optimal in Optimal Control refers to the best weighting of a set of costs associated with particular movements. In the context of speaking, costs might be associated with speaking effort or being misunderstood. A talker evaluating how to articulate the upcoming speech unit plans a

[3]Computationally, this mapping is an inverse Jacobian.

sequence of movements that minimizes these competing costs. While executing those movements, the system runs an efference copy of the motor commands through an internal model of the vocal tract, simulating production of the intended unit. The output of the internal model is an expectation for what the executed articulatory movements ought to sound and feel like. Immediately after execution, SFC compares observed feedback from multiple sources to the expected feedback, ensuring that articulation is progressing as intended. When there is a significant mismatch, the next movement adjusts the articulatory plan in a similarly cost-minimizing way. This is a flexible and powerful way of expressing the problem of speech motor control. Because it is so new, however, many parts of the model remain unspecified, most crucially the interpretation of the control law and the components of the cost function.

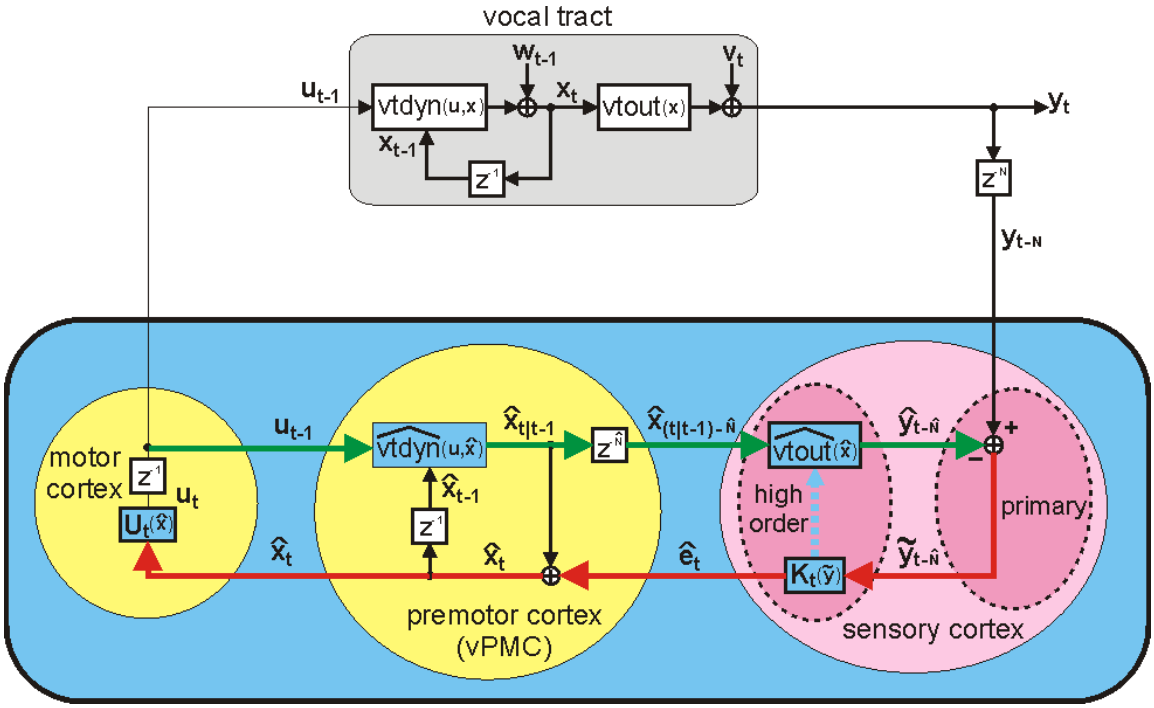


Figure 2.3: Current schematic of the State Feedback Control (SFC) model. From Houde & Nagarajan, under review.

Earlier models of speech motor control cannot fit into this framework. Equilibrium Point models are based on the observation that every innervated muscle has a position that is easiest to maintain, its *equilibrium point*. Groups of innervated muscles likewise have equilibrium points defined by low-energy positions and joint angles. Equilibrium points can define articulation during speech in the following way: suppose that a talker intends to produce [bip]. Each phoneme in that syllable is associated with an equilibrium position, and the syllable is produced by moving

each articulator from its current position toward the equilibrium point of the next phoneme. If the syllable is articulated fast enough, the articulators may not have enough time to reach the [i] equilibrium point before they have to move to the [p] equilibrium point. Incomplete equilibrium point trajectories can explain the formant undershoot inherent to coarticulation. However, this model is not considered further in this dissertation because it fails to account for responses to perturbations. As noted earlier, subjects whose lower lips are momentarily depressed with a paddle while producing [aba] compensate by moving their upper lips downward. This response could be explained within an equilibrium point model by specifying that the upper lip and lower lip have different amounts of resistance to movement and must act as a unit. However, such a model cannot simultaneously predict that subjects will move the upper lip in response to lower lip depression during [aba], and that subjects will not move the upper lip in response to lower lip depression during [afa] (Shaiman & Gracco, 2002); in this model, either the upper lip compensates for lower lip depression or it does not. In other words, Equilibrium Point models cannot accommodate task-dependence. Because compensation in the DIVA and SFC models arises from task-specific motor programs, those two models do not have this problem.

In contrast to these motor-based models, which seek to explain responses to perturbation, Articulatory Phonology is a theory of speech production designed to explain linguistic facts about coarticulation, reduction, and speech errors. Articulatory Phonology relies on a gestural theory of speech production (Browman & Goldstein, 1992), which construe speech as strings of subsyllabic gestures. Though articulatory programs or scores can be stored for larger units, those units are transparently decomposable into their component gestures. Gestures are, therefore, neither phoneme-sized nor syllable-sized: in this framework, the phoneme /p/ is the concatenation of a lip-closure gesture and a laryngeal gesture. Thinking about sounds in terms of gestures has the advantage of accounting for certain coarticulatory effects. For example, prenasalization of the /æ/ in /kæn/ happens because the velic lowering gesture happens before the tongue blade raising gesture, which forms the /n/ closure in the oral cavity. The earlier nasalization is simply a property of *can*'s articulatory score and a property of the *articulatory phonology* of English. Essentially, Browman and Goldstein start from the premise that phonemes cannot simply be concatenated to produce fluent speech, and set about arguing that gestures can be concatenated in this way. Gestural theories also posit that gestures are equally transparent in every language. A dynamical systems approach is taken to coordinate these gestures into articulatory movements (Saltzman & Munhall, 1989). This framework models a "constellation" of gestures representing an articulatory score as coupled springs. The coupling, and therefore the relative timing of the gestures, is set by learned *phasing relations* and *bonding strength*.

Gestural theories of speech production are not as different from models like DIVA as they seem: although DIVA assumes that the acoustic signal is primary, and articulatory phonology assumes that vocal tract constrictions are primary, both models

must perceive speech by converting the incoming acoustic signal to articulatory coordinates. The difference is that a gestural theory would transform the incoming acoustic signal into tract variables specifying the current state of the vocal tract's constrictions, and an acoustic theory would convert the incoming acoustic signal directly to articulatory coordinates.

Psycholinguistic models of speech production

Psycholinguistic models deal with three stages of speech production: lexical selection, phonological planning, and articulatory implementation. Competing models differ mostly in the degree to which they allow interaction between stages of speech production. One of these models represents utterance planning using a multilevel hierarchy, with levels for semantics, the word's form, or lemma, and the word's phonological structure (Dell, 1986, 1988). In this model, each level in the hierarchy contains many nodes that describe its units. For example, a node on the semantic level might represent the concept of ANIMAL, and a node on the lemma level might represent the word DOG. Speakers plan utterances by choosing the semantic content of their message, activating the associated semantic nodes. Activation then spreads downward through connected nodes on lower levels of the hierarchy in a *feedforward* way and also *feeds back* up to higher levels. Once a word is selected and passed to the phonological level, it is broken into syllabic frames. Phonemes cannot be accessed independently of these frames. The model was built to explain facts about speech errors.

Levelt and colleagues (1999) also believe that sentences are planned through activation of nodes in hierarchical stages of planning. Their model differs from Dell's in part because in Levelt et al.'s model, words are retrieved from a mental store along with a phonological code. This code comes with stress rules, but is not presyllabified. For example, the word 'phoneme' would be retrieved as *f-o-n-i-m*. Levelt advocates retrieval of English words as a list of phonemes rather than list of syllables in large part because English undergoes resyllabification. That is, the two words 'phoneme' and 'phonemic' seem to share something phonologically, but a purely syllabic model would retrieve the two words as *fo . nim*, and *fo . 'ni . mɪk*, and in a purely syllabic framework, the two words would not have a single unit in common.

Once the phonological string is retrieved from the mental store, a process called *prosodification* makes decisions about syllable breaks based on the context, and the newly syllabified *phonological word* is passed to phonetic encoding. Phonetic encoding entails lookup of syllables in a mental syllabary containing fairly abstract 'articulatory scores' as described by Articulatory Phonology (Browman & Goldstein, 1992). This system describes speech production as a sequence of constriction goals at a number of possible places in the vocal tract. For example, the first syllable of 'phoneme' might be expressed as a [critical] constriction at the [dental] location followed by a release of those constrictions and a [wide] constriction at the pharynx. These

schematic plans are subsequently modulated for force and duration, which allows the computation of more detailed instructions (e.g., round lips) which are implemented through a dynamical system coordinating the gestural timing and outputted through an "articulatory network" for speaking (Saltzman & Munhall, 1989; Goldstein & Fowler, 2003). Because multiple articulations are used for the same vowel both in compensatory speech (Lindblom et al., 1979) and in normal speech (Maeda, 1991), targets are better phrased in terms of an area function in the vocal tract rather than a set of spatial targets or trajectories for articulators.

There are two main differences between these two types of models. One major difference is that feedback between levels is present in Dell's model, but not in Levelt et al's model. This difference has been a source of voluble debate in the literature, but is outside the scope of this chapter and will not be discussed here. The other major difference is in their description of the units that are used in articulatory planning. Dell believes that words are spelled out in phonemes which are already tagged for syllable position. In contrast, Levelt & Wheeldon (1994) claim that all languages have a level of phonetic encoding containing a mental syllabary on top of their phonological encoding. This syllabary is sensitive to frequency. Thus Dell's model implicates the syllable as the basic planning unit, while Levelt's model ultimately favors a phonemic spellout, followed by a syllable phrased as an articulatory score.

Computational implementations exist for both types of models. The first of the implementable models was TRACE, a feedforward and feedback model (McClelland & Elman, 1986). The corresponding feedforward model (which does not allow interaction) is called WEAVER++ (Roelofs, 1997; Levelt et al., 1999). DIVA and SFC differ from implementations of the Levelt model in that they picks up where cognitive models leave off, specifying how a planned unit gets translated into accurate, continuous speech. However, it does cover the same domain as Articulatory Phonology. The crucial difference between DIVA, SFC, and Articulatory Phonology is that Articulatory Phonology does not allow for signal acoustics to be monitored during production, whereas DIVA requires it.

Psycholinguistic theories of perception make use of the same hierarchy of linguistic representations. An early, influential such model was Cohort Theory (Marslen-Wilson & Welsh, 1978). This model takes in features extracted from windows of 100-150 ms. Based on the initial window, a set of possible word candidates is generated, and as input continues to stream in, candidates are removed from the list. For example, 'beetle' and 'beaker' are initially competitors, but at the onset of the second syllable, they cease to have the same set of features and the ambiguity is resolved. The COHORT model has been criticized for its purely sequential processing of information. Because the set of possible words is generated only once, when the first bit of the word is heard, it is impossible to select the correct word if the first phoneme is misarticulated.

More recent models of word use a variety of methods to generate a list of possible words from speech input. One method is to create a list of compatible words

containing the first few phonemes, as the original COHORT model does. Another is to generate probabilities of hearing particular phonemes, as the MERGE (Norris, McQueen, & Cutler, 2000) or SHORTLIST (Norris, 1994) model does. A third is to generate word probabilities from features, as TRACE does. Depending on how one sets up one's model of word recognition, different subword representations are generated along the way. For example, a model like TRACE (McClelland & Elman, 1986), which takes sets of features as input, has explicit memory for the string of features inherent in all of a word's phonemes. A model like MERGE, which generates an optimal phoneme parse based on all of the acoustic input received thus far, has explicit memory for the string of phonemes inherent in a word. A model based on articulatory phonology could take gestures as input (or takes spectra as input and maps those acoustics to gestures); this sort of model requires a memory for gestures, but not for phonemes or features. So many models exist because it is so far impossible to account for subtle priming and inhibition effects, along with observed behavior in speech errors and language disorders, within a single model.

On balance, the behavioral evidence suggests that speech plans are hierarchically organized, and that speakers and listeners have access to multiple representations of the message to be sent or received. Whether one or several of these levels are shared with speech motor control systems during articulation and self monitoring is an open question.

Merging psycholinguistics with neuroscience

On one hand, psycholinguistic and speech motor control models are not so much incompatible as they specify two different aspects of speech production. Psycholinguistic models cover the territory from semantic formulation to lexical selection and phonological lookup. Speech motor control models take in motor plans, after semantic formulation and lexical lookup, and specify how those plans lead to motor programs and articulation. Because both message formulation and message articulation occur during speech, these two types of models are clearly linked; however, it is not known how they are linked, either cognitively or neurally.

This dissertation investigates to what extent these two types of models are linked. DIVA and SFC models presume that observed and expected feedback are considered with respect to the intended syllable (DIVA) or syllable control law (SFC). They have not considered what effects, if any, phonological and lexical neighbors might have on the timing, magnitude, or quality of compensation. Psycholinguistic models of speech production operate on cognitive entities like semantic nodes and lemmas, without much attention to how these entities might be implemented in the brain. Psycholinguistic models state that feedback between stages of speech production does not incorporate feedback from the articulatory level. Speech errors or altered feedback affecting low-level motor control may have reverberatory effects on higher levels, causing priming of syllables, words, or sentences.

Broadly, there are two ways of thinking about how to combine these two types of models. First, results from experiments in speech motor control should be able to speak to models of speech perception and production, and vice versa. Second, information about the neural correlates of speech production from models of speech motor control can be use to infer neurally plausible explanations for some of the behaviors we observe in psycholinguistic experiments. For example, priming, which is normally explained by activation of some unit of semantic meaning, can be pinned down somewhat more precisely if we view it as similar to the inhibition effects observed in cortical suppression when speaking (Houde, Nagarajan, Sekihara, & Merzenich, 2002). Likewise, speech perception and production could be understood more completely if we also located the brain regions responsible for detecting phonetic mismatches in auditory feedback and processing altered feedback with respect to category boundaries (Niziolek, 2010; Houde, Heinks-Maldonado, & Nagarajan, 2006; Tourville, Reilly, & Guenther, 2008; Guenther, Ghosh, & Tourville, 2006; Niziolek & Guenther, 2009; Tourville & Guenther, 2010; Ghosh, Tourville, & Guenther, 2008).

Merging speech production models in psycholinguistics and neuroscience may help to explain the mechanisms behind other perception-production phenomena as well. Understanding how the articulatory-acoustic map operates on one's own speech can help us understand how the that map works in interpreting incoming speech from others. Deciphering that map will be key to understanding how language is acquired by infants; why articulatory aspects of one's speech influences the way one categorizes non-self-produced speech (Fox, 1982; Bell-Berti, Raphael, Pisoni, & Sawusch, 1979; Perkell et al., 2004); and how hearing non-self-produced speech changes the way one speaks. (Babel, 2009; Nielsen, 2008; Pardo, 2006).

The experiments outlined in Chapters 3 through 6 investigate the influence of a talker's native language on the mapping of perceived speech to produced speech. Chapter 3 introduces experimental methodology and preliminary experiments. Chapter 4 investigates the effect of phonological inventory on this mapping; Chapter 5 investigates the effect of the lexical inventory on this mapping; Chapter 6 investigates the effect of articulatory and acoustic familiarity on this mapping.

The vowels of California English

Because participants recruited for the experiments in this dissertation are primarily from California, the last part of this review illustrates the formant structure of vowels in California English. California English vowels differ noticeably from the vowels of Standard American English. One salient difference is that California English participates in the merger of the low back vowels [ɔ] and [ɑ]. Another is significant [u] fronting, which is especially acute in coronal contexts. Both of these features are shared with other American dialect areas. Less widespread is a tendency toward [o]-fronting, which has been observed in California English since the 1970's (Hall-Lew, 2009, *inter alia*). It seems likely that the vowel dynamics of [o] led to its recent

fronting; in a dialect where [u] is uniformly fronted, an unfronted [o] glides from [o] to a high, back, non-phonemic [u]. That fronting may be phonologized to apply to the [o] offglide, making it more likely that /o/ is realized as [oʊ] or [oʉ]. These differences are reflected in the formants of California English vowels and are easy to spot in Figures 2.4 and 2.5. Figure 2.4 shows the American English vowel space in two regions of the United States. One is taken from the classic Peterson and Barney (Peterson & Barney, 1952) study, which measured vowels produced by speakers in New Jersey, and the other taken from a follow-up study by Hillenbrand et al. (Hillenbrand et al., 1995), which measured vowels produced by speakers in the American Midwest. These studies took their measurements from vowels in citation form *hVd* words. Both graphs are reinterpreted by Hagiwara (1997). Figure 2.5 shows the California English vowel space as recorded by Hagiwara (1997).

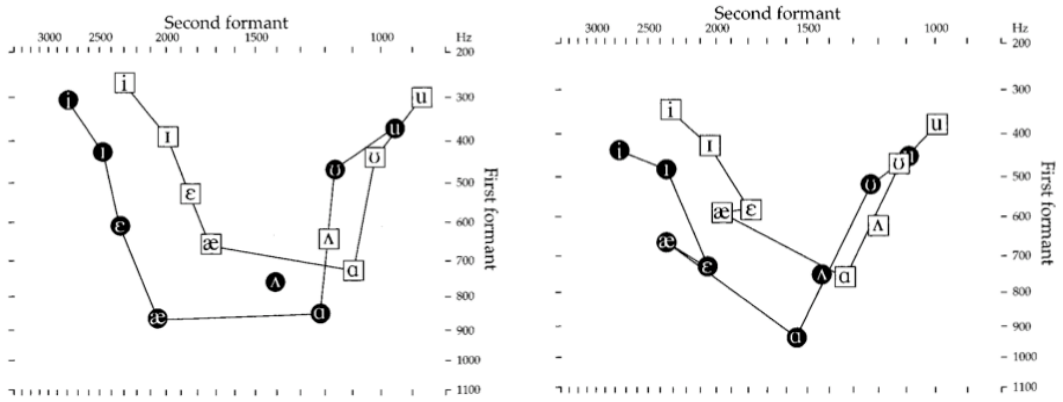


Figure 2.4: Two studies of Standard American English vowel spaces, as reinterpreted by Hagiwara (1997). Men's vowels are in open squares and women's vowels are in filled circles. Left: Peterson & Barney's General American English Vowels. Reprinted with permission. Copyright 1997, Acoustical Society of America. Right: Hillenbrand et al.'s (1995) Standard American English vowels. Reprinted with permission. Copyright 1995, Acoustical Society of America.

Notice that American English vowel formants vary substantially among the three studies. Hagiwara notes that the low vowels /æ/ and /ɑ/ are much higher in the Hillenbrand et al. study than in the Peterson & Barney study because the Hillenbrand et al. subjects spoke a Midwestern dialect that underwent the Northern Cities vowel shift, which involved raising of low vowels. The California English speakers had an /ɑ/ more typical of the Peterson & Barney vowel, but California /æ/ is far further back than the Northern Cities or 1950's American /æ/, and California /u/ is far further front than the Northern Cities or 1950's American /u/. Because the absolute values of formant measurements vary from speaker to speaker, it is instructive to see these vowels plotted on normalized axes. Figure 2.6 shows two methods of doing so. The

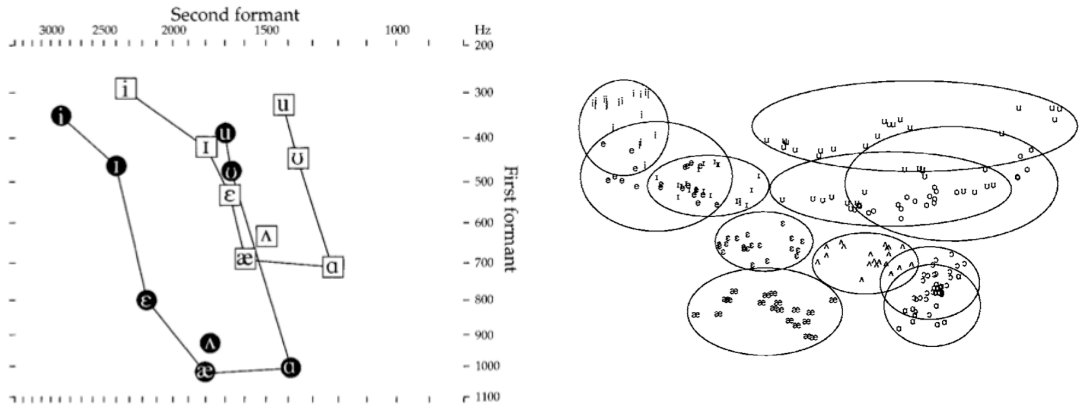


Figure 2.5: Two studies of California English vowel spaces. Left: Data collected by Hagiwara (1997). Men's vowels are in open circles, and women's vowels are in filled circles. Right: Formants from Clopper et al. (2005) Western male vowels, for comparison (all tokens are shown with an ellipse, hand-drawn by the authors, surrounding each cluster). Reprinted with permission. Copyright 2005, Acoustical Society of America.

left graph is a depiction of "Western" vowels from Clopper et al. (Clopper, Pisoni, & de Jong, 2005). The right graph shows the vowels from the Hagiwara study recast in CLIH-normalized axes. The Hagiwara vowels were extracted from monosyllabic words representing 3 consonant contexts. Contexts included *bVt*, *tVk*, and *hVd*. All words were embedded in the frame sentence "Cite CVC twice." The Clopper et al. data were elicited through monosyllabic *hVd* words (with the exception of *frogs* and *logs*).

Figure 2.6 shows that the distribution of vowels and overall shape of the vowel space is quite similar across studies of California English vowels, and the shape of the normalized vowel space squares quite well with the unnormalized data.

The studies in this dissertation were tailored to the composition of California English vowel space. The vowel stimuli in this dissertation were chosen because they were maximally monophthongal in California English. On the basis of the results of this dissertation, other dialects are expected to respond differently to the feedback manipulations practiced here. Cross-dialectal and cross-linguistic work is essential to verify the generality of the results documented in this set of studies.

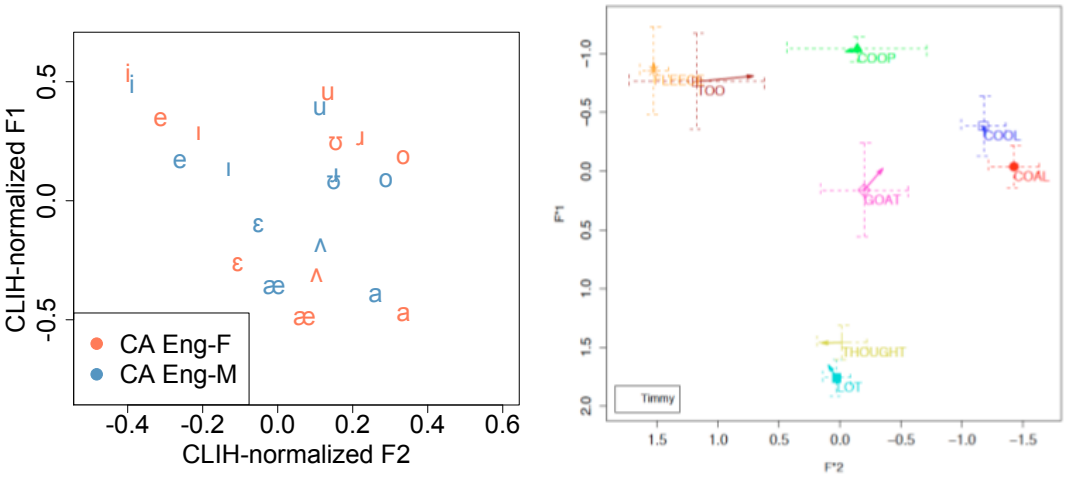


Figure 2.6: Two studies of California English vowel spaces. Left: Normalized formants from Hagiwara 1997's male and female California English speakers. Reprinted with permission. Copyright 1997, Acoustical Society of America. Male speaker averages are in blue and female speaker averages are in orange. Right: Vowel formants collected by Hall-Lew (2009). All tokens are graphed on normalized fomant axes.

Chapter 3

Methods

Acoustic feedback is altered in these experiments with a custom feedback alteration device. The device and the algorithms it uses to find and shift peaks in acoustic spectra are described in the first half of the chapter. The second half explains the reasoning behind the studies described in the remainder of the dissertation.

The Feedback Alteration Device

Subjects in all experiments wear a AKG HSC-271 Professional headset that is connected to a feedback alteration device. When a subject talks, his[1] speech is picked up by the microphone and routed through a phantom-powered preamp and Delta-44 sound card into a computer running feedback alteration software designed by John Houde . The feedback alteration software analyzes the incoming speech and alters formants as programmed, then re-synthesizes the speech and sends it to the headset's earphones. When the feedback alteration device (FAD) does not alter incoming speech, talkers hear their voice played back to them veridically in real time, but when the FAD is set to alter F1, F2, or F3, the talker hears an altered version of his own voice in real time.

The FAD works by analyzing, then re-synthesizing speech input using McAulay-Quatieri synthesis (Quatieri, 2002; Quatieri & McAulay, 1992; McAulay & Quatieri, 1991; Quatieri & McAulay, 1986). The signal is analyzed in 36ms windows, which are updated every 3ms by adding a new frame and deleting the oldest frame. Because the sampling rate is 11,025 KHz, each 3ms frame contains 32 samples. Incoming speech is analyzed with a window shape that gives more weight to more recent information in order to reduce the perceived delay in re-synthesis.

[1]Descriptions of equipment and experiments use masculine pronouns because all subjects in these studies are male. This constraint on the subject population was necessary for successful operation of the online signal processing algorithms. Other types of feedback alteration devices are able to manipulate women's voices, and, to my knowledge, no studies using these other devices find gender-stratified responses.

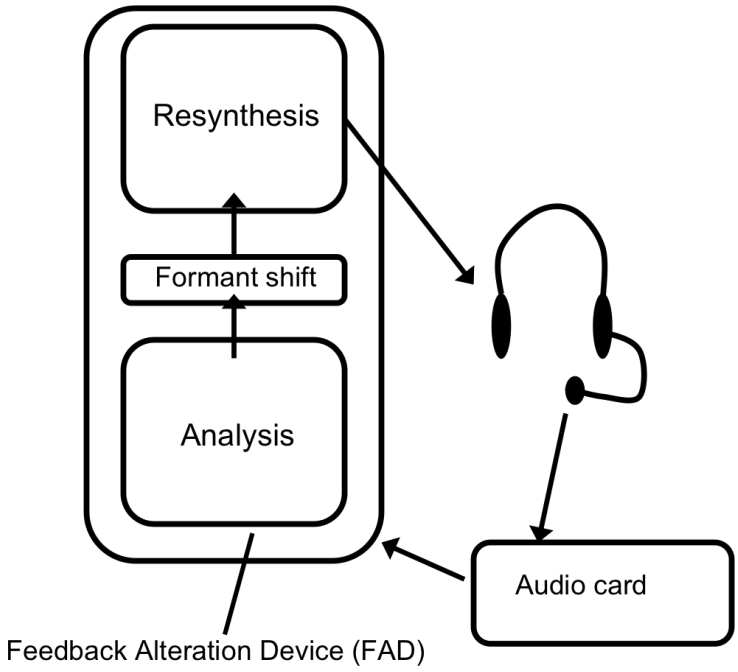


Figure 3.1: Schematic of Experimental Setup. Subjects speak into the microphone portion of a headset. Their speech is analyzed, then re-synthesized (and shifted, if necessary) and fed back into the headset's earphones.

Analysis begins by finding the pitch and spectral envelope in the current window. To find pitch, the FAD computes the discrete cosine transform of the magnitude spectrum. The pitch is the highest peak falling between pre-set pitch bounds in the resulting spectrum. To find the spectral envelope, the shape that a sheet would take if it were draped over the frequency spectrum, the current spectrum is windowed, and all peaks below a threshold magnitude are removed. This magnitude can be changed online. The peaks that remain in the spectral envelope after smoothing are the formants of the spectrum. The formant finding process is illustrated in Figure 3.2 using the vowel /u/ from 'food'. As discussed in Chapter 2, this vowel is substantially different from midwestern American /u/. Peterson and Barney (1952) find that /u/ has an F1 of 300Hz and an F2 of 870 Hz; the /u/ here has an F1 of 301Hz and an F2 of 1924Hz. However, this F2 value is not atypical for California English, as noted in (Hagiwara, 1997).

Two issues arise when using peak picking to find formants. First, it is sometimes the case that false formants are found, or that real formants are missed. To mitigate this possibility, the amount of smoothing can be adjusted before each experimental session. Increasing the amount of smoothing finds fewer peaks, and reducing the

amount of smoothing finds more peaks. This process is analogous to adjusting LPC order to fit a particular person's voice. Second, a talker's pitch appears as a large peak in the spectrum and is sometimes mistaken for the first formant. The FAD deals with this problem by applying a high pass filter to the spectrum before smoothing, removing that first spectral peak from the analysis. The filter's cutoff frequency can be adjusted before each experimental session. In spite of these safeguards, peaks do not always correspond to formants, particularly in women's voices, where it is difficult to set an appropriately high filter cutoff and to smooth away low-frequency false formants while finding higher formants. For this reason, all participants in these experiments were men.

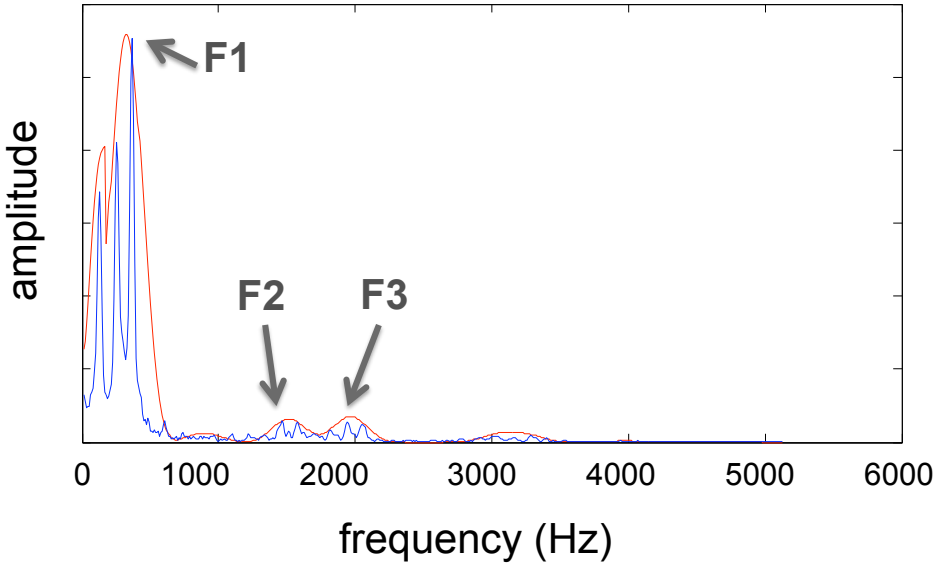


Figure 3.2: Fourier transform of mid-vowel frame from 'food'. The smoothed spectral envelope is overlaid on the spectrum. As marked, formants are peaks in the smoothed spectrum. Formant shifting simply requires changing the value of the formant of interest. The peaks calculated in this particular spectral slice are: F1 = 301 Hz; F2 = 1924 Hz; F3 = 2367 Hz.

The pitch, formants, and frame energy estimated from this analysis are used to re-synthesize the sound. Because the formants are estimated separately, it is trivial to shift formants during the experimental condition. To shift formants, the FAD simply adds or subtracts the amount of formant shift from the relevant formant before re-synthesis. For example, if the incoming formants were analyzed to have F1=300Hz, F2=1924Hz, and F3=2367Hz, and the experiment shifts F2 by -200 Hz, re-synthesis proceeds with the same pitch and frame energy, but with formants F1=300, F2=1724Hz, and F3=2367Hz. Harmonics are calculated from the spectral

envelope, frame energy, and pitch, and a sinusoid is created at each harmonic. The FAD keeps track of the phase of each sinusoid at each frame so that the same sine wave is continuous between frames. Another way of thinking about F1 and F2 shifts is as a change in the amplitude of the harmonics. Harmonics in the position that the formant used to be in are attenuated, and harmonics near the shifted formant are amplified. The sinusoids are summed and outputted to the headphones as the next 3ms frame is inputted. This double-buffering scheme can theoretically recapitulate speech with a 6ms delay (3 ms for frame collection and 3 ms for analysis), but additional buffers in the sound card increase the effective delay to about 12ms, as measured with an oscilloscope. The formant shifting process is illustrated in Figures 3.3 and 3.4. Although altered feedback does not eliminate sound received through bone conduction, and the perception of F1 can be influenced by bone conduction (Pörschmann, 2000), sound played through the headphones was played loudly to mask its effects.

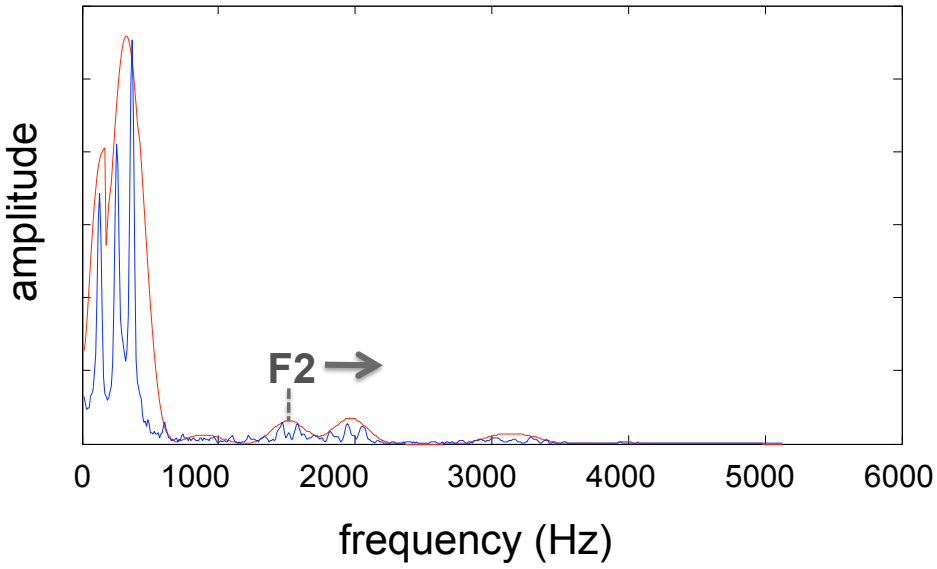


Figure 3.3: Fourier transform of mid-vowel frame from 'food' with direction of shift marked.

The FAD is controlled by a suite of customized MATLAB programs. These programs set values for the smoothing filter, the cutoff for the high pass filter, the frame size, and a host of other online-customizable values that coordinate the timing of the display and details of the signal processing. They also set experimental parameters such as the total number of trials and the formant shift at each trial.

This method of analysis and re-synthesis is different from methods employed in other labs that alter auditory feedback. The most common method of formant alteration uses LPC (linear predictive coding) analysis to find formants. Both the

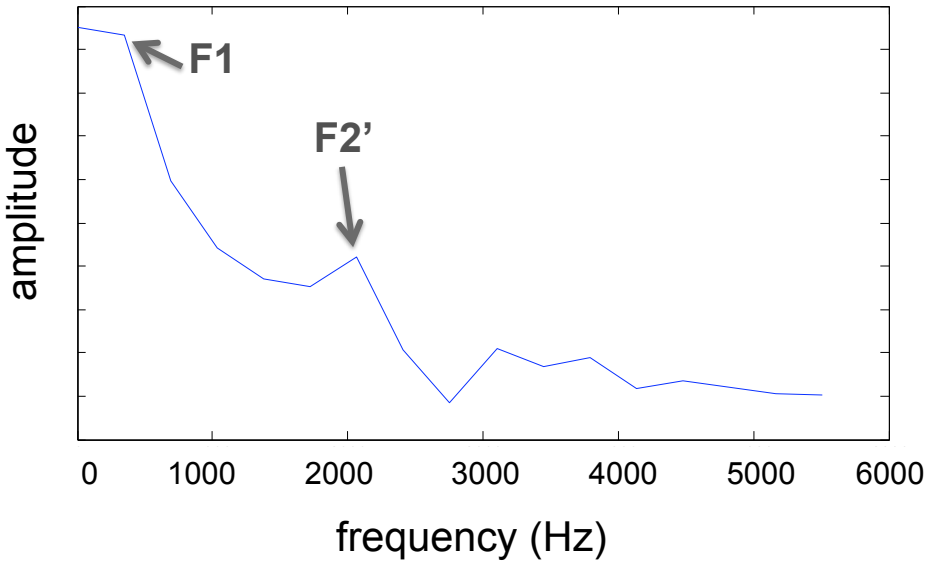


Figure 3.4: Outbuffer from mid-vowel frame of 'food'. The second formant has been shifted by 300 Hz such that it now lies on top of F3. Here, one frame from the shifted word (at 0.57 sec.) is shown.

FAD and LPC divide a waveform into small frames during which the signal is assumed to be roughly periodic and constant. But instead of smoothing a spectral envelope and finding the resulting peaks in the signal, LPC works by predicting the current sample in a frame from the samples that came before it. More specifically, it treats each sample as a linear combination of previous samples, so that each sample can be expressed as a series of coefficients, a_n to a_{n-M}, multiplied by the previous samples x_n to x_{n-M}. Optimal coefficients are found by minimizing the squared error between the predicted signal and the actual signal. These coefficients describe a filter that, when convolved with white noise, re-synthesizes a smoothed version of the analyzed frame. In LPC analysis, vowel formants are locations where the frequency response of the filter reaches a maximum. LPC is considered an *all-pole* model of speech because there are no locations where the frequency response of the filter is zero. To shift formants, LPC-based methods use a notch filter to reduce the amplitude of the spectrum near the produced formant, and a bandpass filter to increase the amplitude of the spectrum near the shifted formant. Unlike the analysis method used by the FAD, LPC does not use peak finding, and smoothing is implicit in the number of coefficients; fewer coefficients result in a more smoothed spectrum and more coefficients result in a bumpier spectrum.

Because of the difference in method, it is important to ensure that the FAD equipment produces changes in production for shifts in auditory feedback that ap-

proximate those found in other labs. The following section shows that the FAD yields similar results to those of Purcell & Munhall (2006b) for 3 subjects.

Preliminary Experiment 1: Replication

Three male participants were seated in a soundproof booth in front of the equipment setup described above. Before the experiment began, subjects read a short passage to become acclimatized to hearing their FAD-resynthesized voices through headphones. No recording of speech or speech alteration occurred during this period. Once accustomed to hearing their re-synthesized voices, baseline vowel spaces were collected by asking subjects to repeat a set of 7 *hVd* words displayed on a computer screen for 20 seconds at their own pace[2]. Feedback was not altered during this stage.

During the alteration stage, a MATLAB program displayed a prompt on the computer screen at regular intervals. To disguise the nature of the task, talkers were told that their reaction time would be recorded as they followed the instructions given by the prompt. They were not informed that their speech would be altered, though a full explanation of the study was given at the end of the experiment.

During each trial, the prompt 'Say HEAD now' was displayed on the computer screen for approximately 1 second. The first formant of the talker's voice was altered during the full trial using the feedback alteration method described above. Post-session interviews indicated that subjects did not notice either formant shifts or delays.

Formant alteration proceeded in four phases:

Phase	**# Trials**	**Formant shift**
1	75	0Hz
2	40	0Hz - 200Hz
3	210	200Hz
4	35	0Hz

Initially, formant feedback was left unaltered. After 75 baseline trials, the talker's F1 feedback was slowly raised to 200Hz higher than was actually produced, and remained at the 200Hz maximum shift for 220 trials. Feedback was returned to normal for the last 25 trials.

As talkers heard their F1 increase from trial to trial, they began to produce /ɛ/ with a lower F1 (so that their vowels sounded more like /ɪ/). That is, they *compensated* for the change in auditory feedback. The time course of this effect for a representative subject is illustrated in Figure 3.5, along with results from Purcell & Munhall's (2006b) experiment. The "heard" formants shown with open circles were calculated by adding the amount of formant shift to the formant that the subject produced. The comparison shows that the results of this replication match the main features of the Purcell & Munhall result.

[2]hVd words included /hid/, /hɪd/, /hɛd/, /hæd/, /hɑd/, /ho^{w}d/, /hud/.

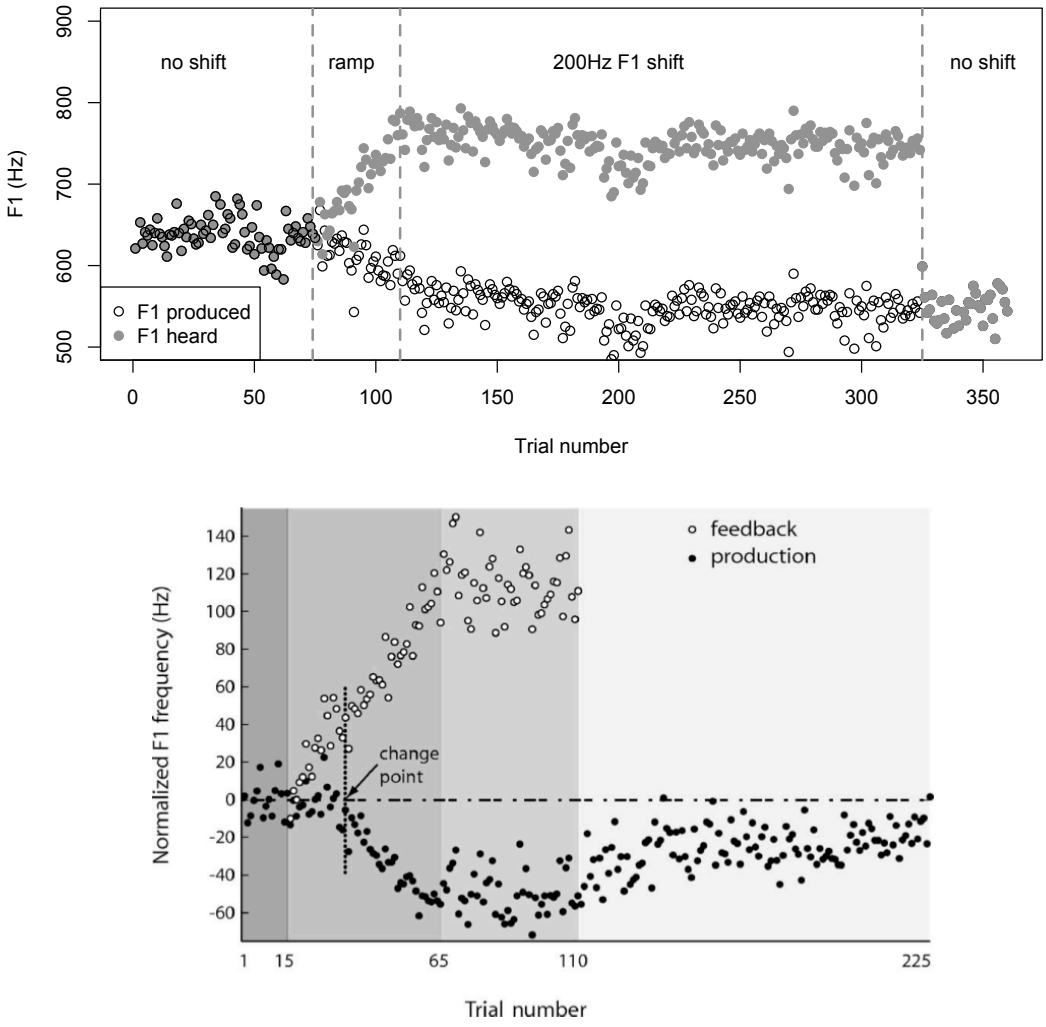


Figure 3.5: Top: Change in F1 feedback and F1 production in /ɛ/ over the course of the experiment for a representative subject. Gray filled circles mark the F1 values that the subject heard at each trial. Open circles mark the F1 that the subject produced at each trial. Thus, each gray circle/open circle pair represents one trial. Bottom: Change in F1 feedback and F1 production in /ɛ/ over the course of the experiment, averaged across subjects from Purcell & Munhall, 2006. Reprinted with permission. Copyright 2006, Acoustical Society of America. Open circles indicate the F1 values that the subject heard at each trial, and black filled circles mark the F1 that the subject produced at each trial.

Experimental procedure

There is no set standard for formant shift size, method, or number of trials; different labs have their own conventions. One of the tasks of this dissertation was to determine whether there was a major effect of experimental design on compensation for formant shifts. Within broad limits, the number of trials and the size of the formant shift did not affect the timing or manner of compensation. The experiments outlined in the remainder of this dissertation are modeled after the Purcell & Munhall, 2006b design, with a short baseline block followed by a formant shift block and a recovery block, but procedural details varied between experiments. Most experiments reported here required 3 nonconsecutive days of testing. The first day was devoted to measuring the subject's baseline vowel space, and the second and third days were devoted to (often complementary) feedback shifts. Several subjects returned on a fourth day to participate in a mock interview (see Chapter 6). A short description of the procedure on each of the three days follows.

Day 1. Subjects were seated in front of a computer screen and the FAD setup described above. Words were displayed on the screen at a rate of about 1 per second. Subjects were instructed to produce those words loudly and clearly. In order to keep subjects from paying an unnatural amount of attention to their auditory feedback, they were not told that the study would analyze their vowels or that their auditory feedback would be manipulated[3]. Recording began as soon as each word was presented and continued for the next 1.1 seconds, which enabled each word to be saved automatically to a separate sound file. After every 15 trials, subjects were permitted to take a silent break and continue when they were ready.

Control sessions contained either 300 or 360 words, depending on the study. They were a monosyllabic mix of words and nonwords from the set $\{b,k,d,h\}\,V\{d,g\}$. Feedback was not altered on Day 1.

Days 2 and 3. The second and third days usually had a 5-block design.

1. Pre-manipulation perception test. Subjects categorized words along a 7-step /hɛd/-/hæd/ continuum. Words were presented in a random order, and each word appeared twice.

2. Pre-manipulation vowel space. Subjects produced a set of 100 $\{h,b,k,d\}\,Vd$ words while wearing the headset. As on Day 1, words appeared on the computer screen at a rate of approximately 1 per second. Re-synthesized auditory feedback was present but not altered during this condition.

3. Manipulation. Subjects produced words that appeared on the computer screen at a rate of approximately one per second. Subjects produced a total of 200

[3]Several subjects who participated in later experiments after having been informed about the formant manipulation still compensated for the feedback shift. This finding is in line with published work on the topic, e.g. Munhall, MacDonald, Byrne, & Johnsrude, 2009.

words during this session and were given the opportunity for a silent break after every 15 trials. Auditory feedback was slowly manipulated over the course of these trials. The feedback manipulation varied with the experiment.

4. Post-manipulation vowel space. Same procedure as pre-manipulation vowel space.

5. Post-manipulation perception test. Same procedure as pre-manipulation perception test.

Analysis

Formant extraction was performed by a series of scripts that make use of Entropic's Speech Processing System (ESPS)/Xwaves, a UNIX-based set of speech analysis utilities installed in the Phonology Lab at the University of California at Berkeley. A perl script ran the ESPS utilities on each sound file, extracting voiced portions of the signal that were greater than 50ms in length and no less than 40% of the maximum signal amplitude, and using a 12-pole LPC analysis to measure their formants. This choice was appropriate for the data files collected in this experiment; each monosyllabic word was saved to a different sound file, and every file contained exactly one vowel. Analyzing the vowel formants post-hoc using LPC even though the formants were already estimated within the FAD had the added advantage of double-checking the FAD's re-synthesis. If the FAD were to mis-estimate formants or to miss important non-formant details in the signal, the resynthesized speech would be poorly reconstructed. If LPC could find the intended formants in the reconstructed signal, it was deemed a successful trial. Some tokens were analyzed by hand if LPC failed to find appropriate formant measurements, but most were simply removed from the analysis. Poor LPC tracking was a problem in less than 5% of tokens. Tokens greater than 2.5 standard deviations from the mean were considered outliers and not considered in the analysis.

Even assuming accurate formant measurement over the course of an experiment, measuring and interpreting *compensation* is not straightforward. Following is an overview of some of the issues encountered in the experiments described in Chapters 5 and 6, and the decisions made to deal with them.

Once such issue is autocorrelation, which causes formants in two consecutive instances of the same word, or even two consecutive words containing the same vowel to be more similar than they would have been had they been recorded several hours or days apart. In other words, autocorrelation yokes successive vowels to each other, limiting the amount of formant change possible from one trial to the next. Autocorrelation was observed in a (2006b) study by Purcell & Munhall as well. They noted that autocorrelation was significant at lag 1 but not at greater lags, meaning that the current trial is dependent on the previous trial but not on trials before the previous

trial. They also found a cross-speaker correlation between the amount of (negative) autocorrelation during no-shift trials and the maximum amount of compensation during trials with maximum shift. They tie these results to force field learning but do not offer an explanation for the effect.

In this experiment, 360 no-shift trials were available from the first day of the experiment, permitting autocorrelation to be measured with greater certainty even in the absence of feedback shift. Although both this experiment and Purcell & Munhall, 2006b found significant inter-subject variability in autocorrelation, our experiments differ in the range of autocorrelation observed. In the current experiment, two subjects had negative correlation at lag 1, indicating that their F1 tended to drift lower over time, and five had positive correlation, indicating that their F1 tended to drift higher over time. Between subjects, autocorrelation varied from -0.545 to 0.557 (mean 0.14). In Purcell & Munhall's study, autocorrelation ranged from -0.73 to 0.04.

Autocorrelation over the course of the entire F1 shift experiment ranged from -0.063 to 0.520 (mean 0.15). And as Purcell & Munhall found, there was a small negative correlation between a subject's autocorrelation during control trials and their mean compensation at the maximum shift. In their study, the correlation was -0.44. In this study, the correlation was -0.31.

Clearly there is something to be explained. Successive trials do not have to be yoked to each other in this way. It is possible that subjects try to recycle pieces of the current motor plan when generating the next one. As a practical matter, this limit implies that multiple trials are potentially required to reach full compensation, and experiments with a small number of trials may make compensation erroneously appear partial. To safeguard against this effect, maximum formant shifts were held, in most experiments, for 90 trials, in order for maximum compensation to be reached by midway through the set of trials with the maximum formant shift.

A second concern was which acoustic parameters to measure. Subjects who compensate tend to oppose the change they hear in that, if their voice feedback has a raised F1, they will speak with a lower F1. They will also, however, change their production of F2, and plausibly other components of their speech as well. This is a concern because calculating a subject's change in production requires deciding which dimensions might register a change. If one were to look at changes in F1 production that result from F1 feedback shifts, subjects would appear to have compensated less than they actually did. Understanding which dimensions actually change is also important for understanding processing of auditory information. Subjects who *can* produce an /ɛ/ with a F1 that is 100Hz higher, but instead produce an /ɛ/ with an F1 50 Hz higher and an F2 50 Hz higher, may perceive incoming vowels as a combination of formants rather than as individual formants. To account for compensatory changes in multiple formants, the experiments described in Chapters 4, 5, and 6 measure compensation in both F1 and F2.

A third concern is where to measure the vowel, and how many measurements per vowel to take. It is clear that American English vowels, even monophthongs, exhibit

a good deal of spectral change (Nearey, 1989). It is also clear that spectral change helps to determine consonant context (e.g., Lindblom, 1963; Strange, 1989), and that rate of spectral change can be important to vowel perception (Nearey & Assmann, 1986; Carré et al., 2001). It is not clear that the California English spoken by the subjects in these experiments had the same spectral dynamics as those measured in these other studies of American English. Of most interest are the dynamics inherent in the California English vowels /ɛ/, /ɪ/, and /æ/. To measure the dynamics of these vowels, vowel trajectories were measured in a representative speaker of California English who participated in this experiment. Figure 3.6 shows these vowel trajectories and companion data from Hillenbrand and Nearey (1999). In these graphs, each line segment represents one vowel. Formants at vowel onset are at the unlabeled end of the line segment, and formants at vowel offset are at the labeled end of the line segment. Both studies show some diphthongal behavior in all vowels with the possible exception of /i/. But even for vowels where California English formants are not known to differ from midwestern American English formants, the two have substantially different formant trajectories. For example, Hillenbrand and Nearey's /ɛ/ shows a small decrease in both F1 and F2 over time, equivalent to centralization. By contrast, this California English speaker shows a substantial increase in F1 and no consistent change in F2 over the course of the vowel. The California English speaker's /ɛ/ shows such an increase in F1 that /ɛ/'s offset formants are the same as /æ/'s onset formants. Perhaps surprisingly, the /æ/ vowel has not diverged from midwestern English in response to the /ɛ/-/æ/ overlap; the /æ/ vowel in California English is very similar to the corresponding vowel in midwestern American English. Both show a moderate increase in F1 and a decrease in F2 over the course of the vowel, which essentially pushes /æ/ toward the lower edge of vowel space. Perhaps this movement in /æ/ prevents it from being confused with California English /ɛ/.

Vowel dynamics are a potential issue in explaining differences in compensation between vowels that have different spectrotemporal characteristics. It is possible to incorporate vowel dynamics into one's explanation of asymmetries in compensation for altered auditory feedback for particular vowels: for example, perhaps hearing steady-state /ɛ/ with a lower F1 sounds like steady-state ɪ, but hearing /ɛ/ with a higher F1 does not sound like /æ/ because /æ/ has different temporal characteristics than /ɛ/ does. I try to consider the consequences of spectral changes in vowels in interpretating the results of the experiments presented here.

Preliminary Experiment 2

The goal of this dissertation is to document individual differences in compensation for altered auditory feedback, and use this information to test models of speech motor control. The experiments that form the main contribution of this dissertation were motivated by the following preliminary experiment, which demonstrates that (1) compensation for altered auditory feedback is incomplete, that (2) some of the

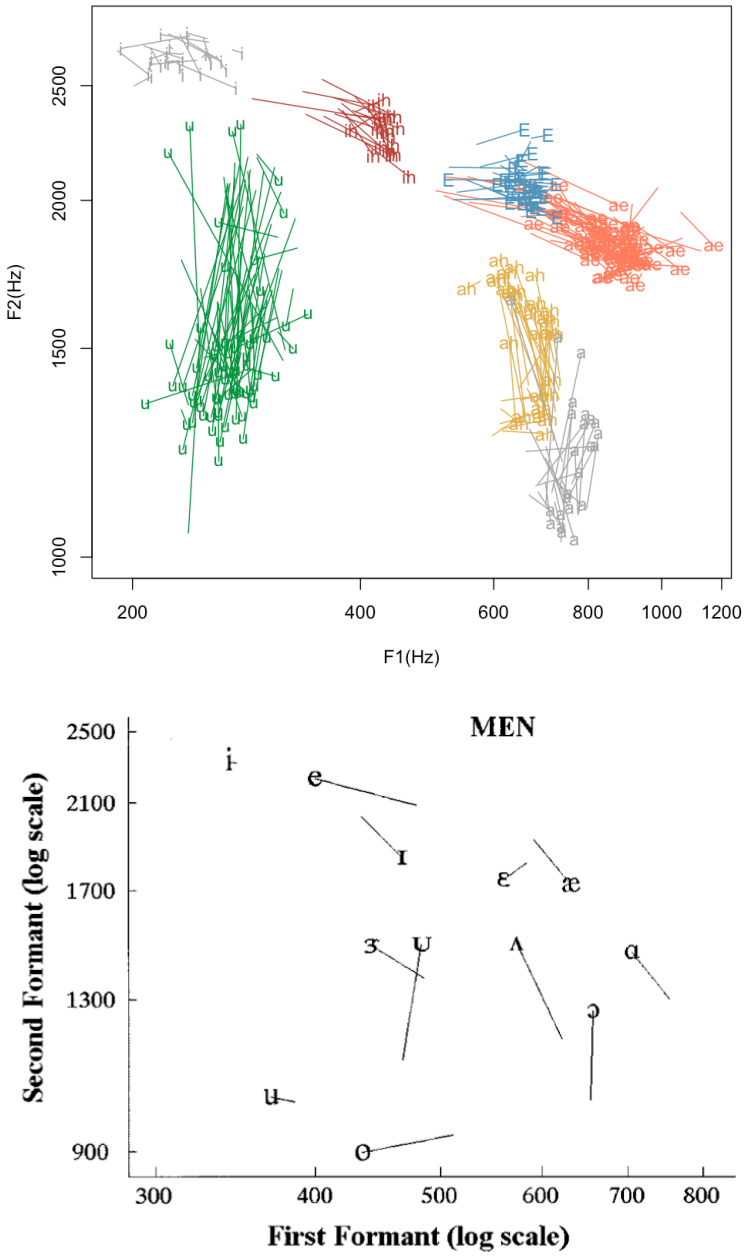


Figure 3.6: Top: Vowel trajectories for a speaker of California English. Each line segment represents one vowel produced by this speaker. Formants at vowel onset are at the unlabeled end of the line segment, and formants at vowel offset are at the labeled end of the line segment. Bottom: Trajectories of midwestern American English vowels, as reported by Hillenbrand and Nearey (1999).

incompleteness in compensation can be explained by vowel targets that incorporate both auditory and somatosensory feedback, and that (3) other characteristics of compensation cannot be explained by multimodal vowel targets alone. It is likely that an individual's physiology, perception, or linguistic organization also affects compensation for altered auditory feedback.

In this experiment, the equipment setup and initial word lists were the same as in Preliminary Experiment 1 except for the alteration phase. In this experiment's alteration phase, formant feedback was slowly raised or lowered to five distinct maximum shifts in a stepwise fashion, resulting in a staircase pattern of formant feedback shifts. Altered feedback remained constant for 20 trials at the top of each "stair". To give subjects adequate time to reach maximum compensation at each feedback shift and to mitigate the effect of autocorrelation between trials, each "stair" in this protocol was 20 trials long. Trials at each stair were treated as if they were exchangeable, and indeed, speakers appeared to compensate by approximately the same amount at each trial on each stair. The progression of feedback shifts during the alteration stage is illustrated in Figure 3.7. There were 7 participants in this experiment, all adult males. Each F1 experiment ramped F1 feedback from baseline (no alteration) to a large maximum alteration of 250 Hz.

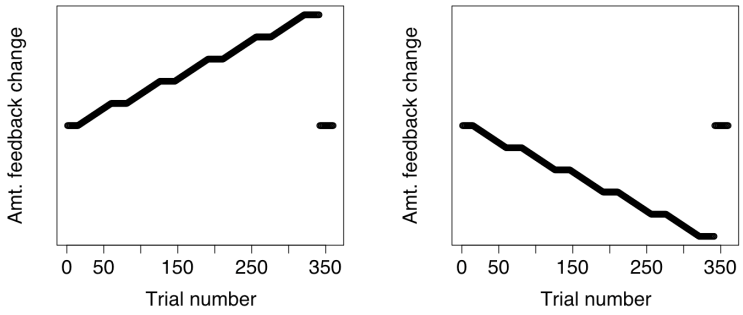


Figure 3.7: Change in feedback over the course of each experiment. There were 360 trials in each experiment. These 360 trials were composed of 6 regions of equal formant alteration ("stairs") connected by ramps of slowly increasing or decreasing feedback alteration.

All subjects in this experiment compensated for the shift in F1 feedback. A typical subject's F1 in /ɛ/ over the course of the experiment is illustrated in Figure 3.8.

The F1 in this talker's /ɛ/ clearly decreased for increasing formant shifts. Relating these raw formant values to percent compensation requires taking into account the baseline F1, which was recovered from the baseline condition of the experiment.

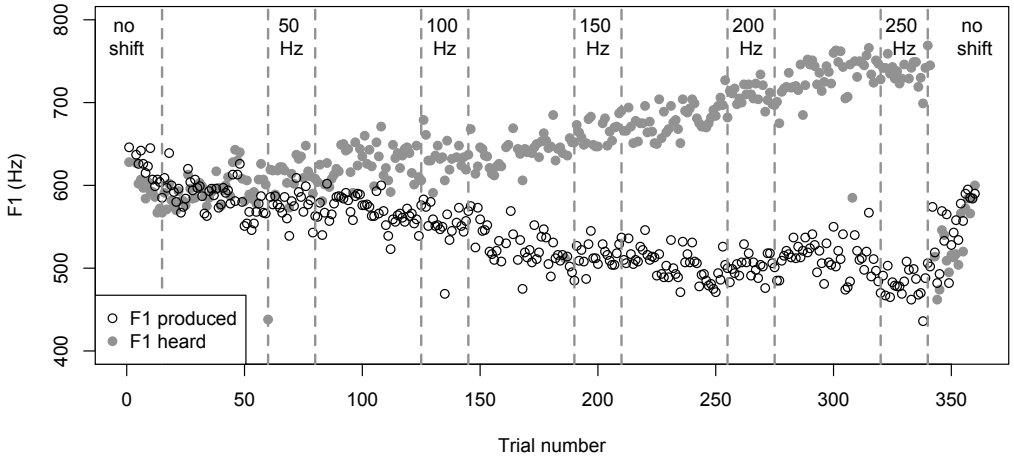


Figure 3.8: F1 from the /ɛ/ in 'head' over the course of the experiment (one typical subject shown here). Open circles mark F1 from the vowels that the subject produced at each trial, and gray filled circles show the altered F1 heard by the subject at each trial. Each gray circle/open circle pair represents one trial.

The baseline condition, during which subjects produced 360 instances of 'head' with no formant shift, catalogued the acoustic variety in /ɛ/ formants typically produced by that subject. A convex hull surrounding the F1 and F2 produced for all of these vowels is shown in Figure 3.9. The standard deviation of these baseline vowel regions is approximately 30 Hz, which is in line with other, similar studies (e.g., Purcell & Munhall, 2006a, *inter alia*). This and all subsequent analyses are performed in Hz and also in Bark, a psychoacoustic scale based on the frequency response of the cochlea (Zwicker & Terhardt, 1980).

Figure 3.9 shows that the first 15 /ɛ/ formants from day 1 (outlined in gray) occupy but a small portion of the /ɛ/ vowel space recorded during the control condition (outlined in black). The means of the gray region and the black region differ because high variance and autocorrelation conspire to make those first 15 trials poor representatives of the subject's true baseline. The first /ɛ/ recorded on Day 2 might lie anywhere within the large vowel region, and the formants recorded during the next 15 trials are influenced by that first vowel's location. Because calculation of percent compensation is dependent on an accurate calculation of the subject's baseline, those first 15 trials were augmented with the additional 360 baseline trials to build a more comprehensive target region for each subject.

Using this augmented baseline, percent compensation was estimated within and across subjects using a mixed-effects linear model. The baseline estimated under this

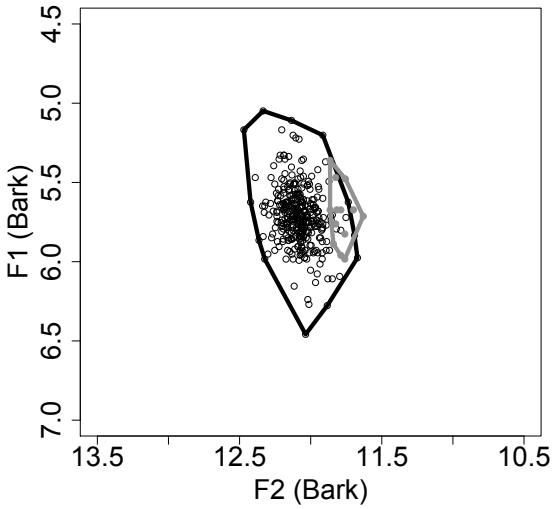


Figure 3.9: A typical subject's baseline region for the /ɛ/ in 'head'. Open circles mark the vowel formants extracted from the 360 vowels produced during the control condition, and the solid black line is the convex hull surrounding them. Gray circles mark the vowel formants produced during unaltered trials of the F1 shift experiment, and the smaller, gray convex hull outlines them.

model is a single point, but that point is estimated using formant measurements that cover the large, control vowel space rather than the small 15-trial baseline from the beginning of a particular session. This method allowed us to use all of a subject's baseline trials when deciding on his baseline during a given session.

The model allows for straightforward estimates of percent compensation along with confidence intervals for those estimates. Two major results arise from this analysis: (1) compensation is almost never complete, and (2) compensation decreases for increased formant shift.

The model describes the F1 produced as a function of a baseline F1, which was permitted to vary by subject, and the formant shift size.

$$\begin{aligned} F1 = \text{baseline}_i \quad &+ \quad \beta_j * \text{shiftsize}_j, \\ \text{where } 1 \leq i \leq \#\text{subjects} \quad &\quad \text{and } 1 \leq j \leq \#\text{shifts} \end{aligned}$$

Subject is included as a random effect in this model. In other words, each subject is assumed to have an idiosyncratic baseline F1, estimated from his 360 baseline vowels. When their feedback is shifted, subjects are assumed to aim to change their production by the same proportion of the feedback shift. The estimated coefficients for the feedback shifts can be interpreted as subjects' mean percent compensation.

In this model, every subject can have his own percent compensation that falls some distance from the mean percent compensation. When subjects do not compensate at all (percent compensation is 0), they produce their baseline F1, and when subjects compensate fully (percent compensation is 100), they produce an F1 that exactly opposes the feedback shift. Ninety-five percent confidence intervals for these estimates were obtained with 20000 iterations of Markov Chain Monte Carlo sampling of model parameters, using the `languageR` package of the statistical software R (see Baayen, Davidson, & Bates, 2008 for a clear account of how these confidence intervals are calculated). Using this model, each subject's baseline and percent compensation was estimated at each of the five formant shift plateaus.

The mean estimates of percent compensation at each of the five plateaus along with their confidence intervals are shown in Figure 6. This figure illustrates the following two trends in the data:

1. Confidence intervals for percent compensation for the smallest shift, 50Hz, do not overlap with confidence intervals for percent compensation for the largest shift, 250Hz.

2. Percent compensation is approximately complete at a formant shift of 50Hz, but is partial for all shifts greater than 50Hz.

Figure 3.10 shows a decrease in percent compensation as the shift size increases, with an appreciable decrease in percent compensation between the 50Hz shift and the 100Hz shift, and a smaller decrease in percent compensation between successive shifts above 100Hz. To test the hypothesis that compensation decreases for increasing formant shifts, sets of two points (c_1, c_2) were drawn from the 50Hz and 250Hz compensation distributions using a Gibbs sampler implemented with WinBUGS and the model described above. Percent compensation at a formant shift of 50Hz was greater than percent compensation at a 250Hz formant shift for 100% of draws. This is strong evidence that compensation is partial and decreasing.

This trend is characteristic of individual subjects as well, as shown by Figure 3.11 below.

Figure 3.11 shows a representative subject's baseline and /ɛ/ formants produced during the experiment. Each of the graphs in Figure 3.12 shows the formants during tokens produced at each of the five F1 feedback shift steps: 50Hz, 100Hz, 150Hz, 200Hz, and 250Hz. The dashed shape in each graph outlines the vowels produced during control trials. The shape outlined by a dark solid line in each graph represents the convex hull of vowels produced when F1 feedback was shifted by the amount shown in the graph title. For example, the dark, solid shape in the leftmost graph outlines the vowels produced during trials with F1 feedback shifted by 50Hz. The gray shape in each graph outlines these vowels after they have been shifted by 50Hz. These are the vowels that subjects heard. If the gray outline is contained within the dashed outline, compensation was complete. As an example, consider a vowel produced with

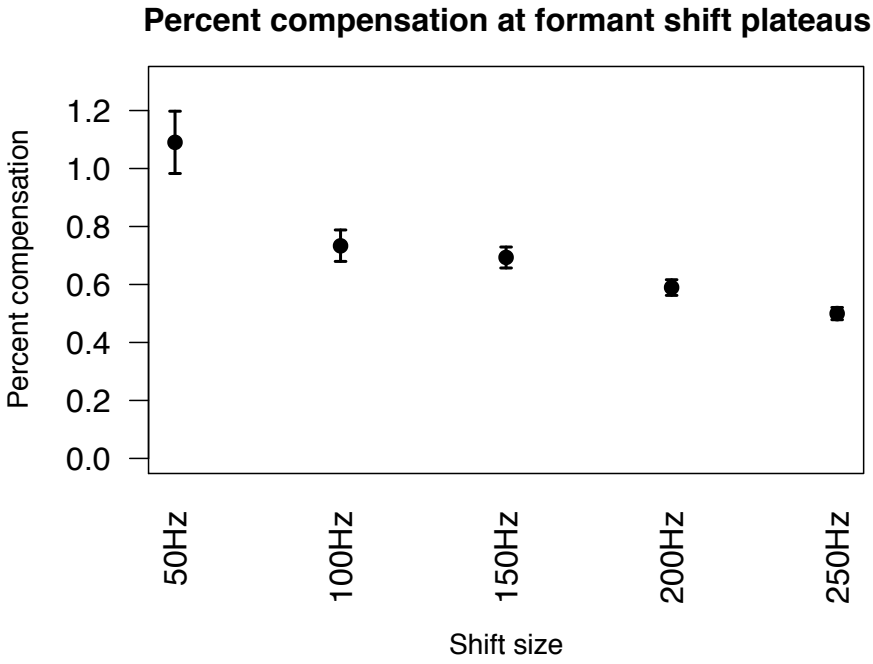


Figure 3.10: Percent compensation, averaged across subjects, at each of the five formant shift plateaus, as estimated from a linear mixed effects model. Error bars mark 95% confidence intervals for each plateau.

an F1 of 5.34 Bark (550Hz) in the leftmost graph. That vowel would fall within the solid black shape. After its 50Hz shift, that vowel would be heard with an F1 of 5.75 Bark (600Hz), which is within the solid gray shape. Notice that in the leftmost graph, the gray shape falls almost completely within the dotted shape, indicating that the shifted vowels that the subject heard were almost all within his baseline region, and that compensation was nearly complete. Compensation is likewise nearly complete for F1 feedback shifts of 100Hz. As the amount of feedback shift increases to 150Hz and beyond, the vowels that the subject hears are no longer within his baseline region and compensation is less and less complete.

An investigation of raw compensation offers a more complete account of the relationship between compensation and formant shift. If subjects were compensating the same absolute amount for each formant shift, percent compensation would appear to be decreasing: a 50 Hz change in production is 100% of a 50 Hz formant shift, but only 25% of a 200 Hz shift. Figure 3.12 demonstrates that this this not the case because raw compensation increases nonlinearly with increasing formant shift. Percent compensation decreases not because the larger feedback shifts have larger denominators in the calculation of percent compensation, but because the increase

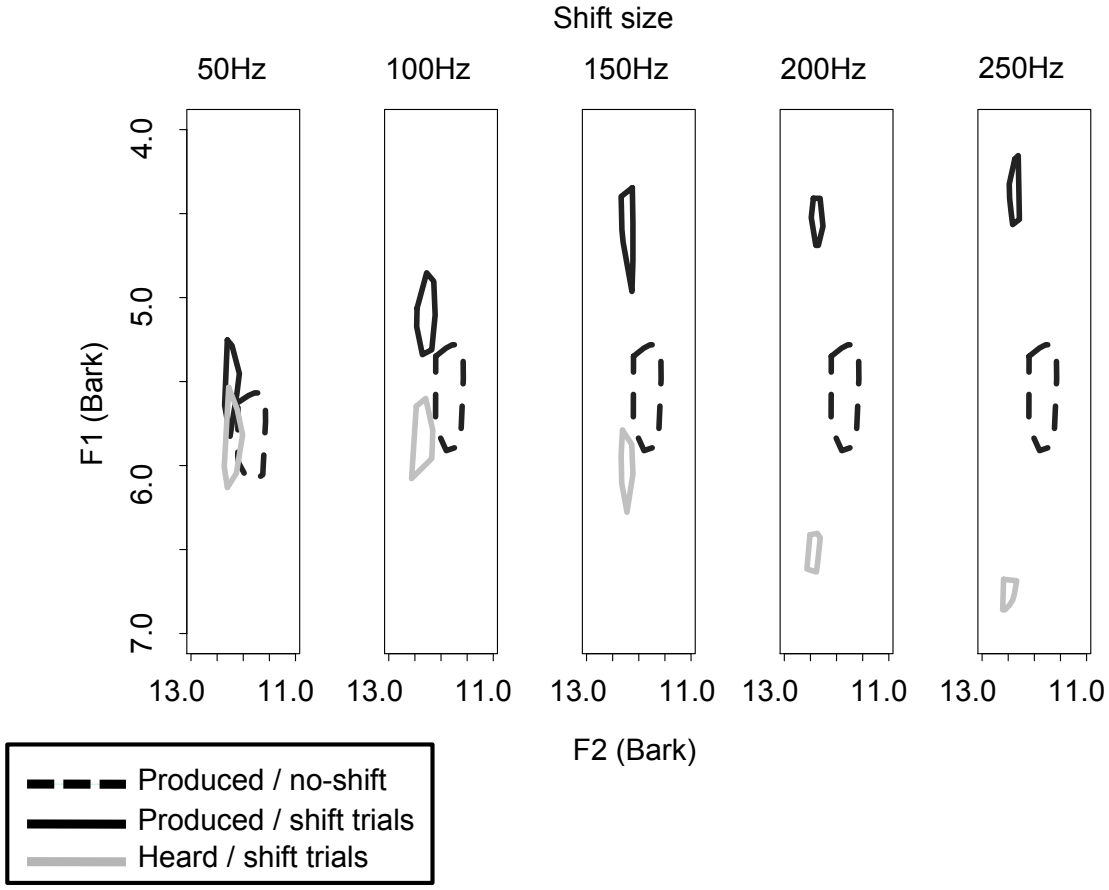


Figure 3.11: Productions of /ɛ/ during Experiment 1 plotted in F1-F2 Bark space against productions of /ɛ/ during control trials. Results for a typical subject are shown. For small feedback shifts, the light gray shape (formants heard as a result of the feedback shift) falls entirely within the dashed shape (the subject's baseline range), indicating that the vowels that the subject heard were all within his baseline region and that compensation was complete. As the amount of feedback shift increases (the dark solid shape), compensation is less and less complete.

in compensation does not nearly keep pace with the increase in feedback shift. Error bars were determined using the same method as Figure 3.10.

Figure 3.12 demonstrates that the amount of additional compensation is smaller at each successive formant shift after 150 Hz. It is possible that, for a sufficiently large formant shift, absolute compensation may approach an asymptote. Indeed, when MacDonald, Goldberg & Munhall (2010) altered F1 and F2 feedback by 350 and 400Hz, they found that compensation approached an asymptote, then decreased at the highest levels of compensation.

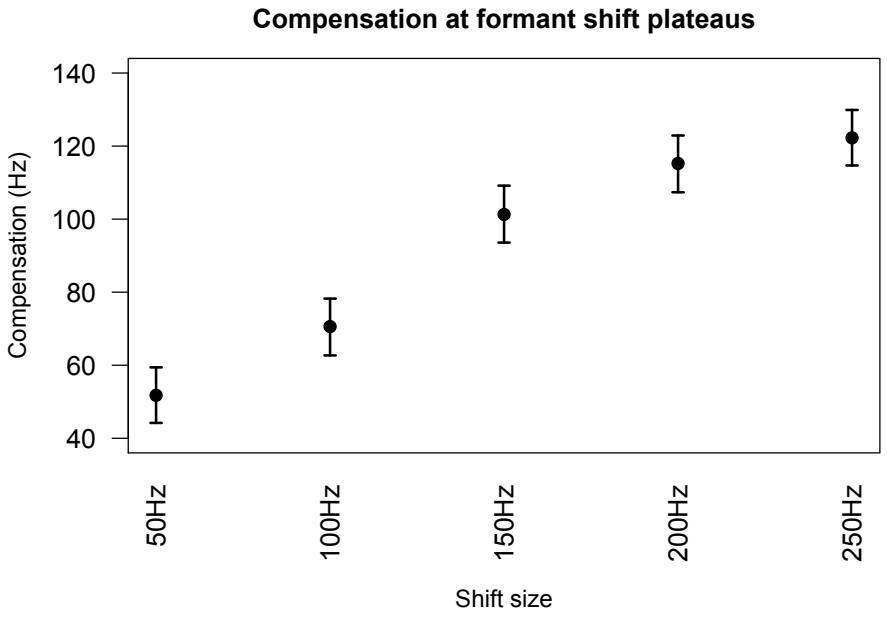


Figure 3.12: Raw compensation in Hertz, averaged across subjects, at each of the five formant shift steps. Error bars mark 95% confidence intervals for each plateau.

This experiment used a stepwise feedback alteration design and a novel method of quantifying baseline vowel regions to measure compensation for feedback alterations of five sizes. Overall, compensation decreased monotonically as the formant shift increased. Specifically, compensation was approximately complete for small shifts in auditory feedback and partial for large shifts in auditory feedback. But there were a number of issues associated with the data and the experimental design.

One major concern is that, in spite of robust group effects in compensation experiments, including this one, there are large individual differences in compensation. In all feedback perturbation experiments, whether the perturbation is auditory or somatosensory, it is common for one-fourth to one-third of subjects to fail to compensate. In spite of altered feedback, these non-compensators make no change in their vowel production at all. Chapter 6 develops a model of compensation based on language production habits to explore whether a subject's native language determines how he will compensate.

Another concern is that the experimental design is highly artificial. Although subjects are told that their speech will be played for future subjects, they are alone in a sound booth during the study itself. The experiment described in Chapter 4 asks whether compensation for auditory feedback differs between words and nonwords, and whether this difference is affected by having to use these words communicatively for a conversational partner.

Finally, we observed that, for large feedback shifts, the re-synthesized vowel fell outside of the subject's natural vowel region. There is some evidence that different brain regions are recruited to deal with feedback perturbations that fall outside of the intended vowel region (Niziolek & Guenther, 2009). There is also some possibility that our baseline measure inflated percent compensation within vowel boundaries, since measurement is less precise on such a small scale, and because the baseline measure was a composite of that day's trials and trials from the control day.

Rationale for the following three studies

Preliminary experiment 2 showed that talkers do not compensate completely for altered auditory feedback. One plausible explanation, suggested in the discussion of that experiment, is that incompleteness in compensation arises from a difference in accuracy between acoustic and somatosensory feedback. Although acoustic feedback is altered systematically over the course of the experiment, somatosensory feedback is not. It is possible that the speech motor control system seeks to split the difference between the two, resulting in partial correction of altered auditory feedback. This explanation presupposes that the calculation of the mismatch between observed and auditory feedback, and the method of deciding how to compensate for it, does not receive top-down influence from the lexicon or phonological inventory. In the speech motor control literature, expected auditory feedback has been proposed to come from one of two sources: (1) an internal estimate of the speaker's current articulatory state, as generated by passing the current set of motor commands through an internal model (Körding & Wolpert, 2004; Bhushan & Shadmehr, 1999), or (2) an acoustic region associated with the intended phoneme or syllable accessed from memory. Both propose that corrections for mismatches are sensitive to articulatory plans. If the articulatory plans are the size of a linguistic object (e.g., phonemes or syllables), then these aspects of language will affect compensation for feedback alteration. The remainder of the experiments in this dissertation seek to determine whether there is really no higher-level linguistic influence on low-level auditory feedback.

The first experiment, in Chapter 4, investigates the *word avoidance hypothesis*, which proposes that top-down information from the lexicon influences compensation for altered auditory feedback. There are two conditions in this experiment. Stimuli in the first condition are designed such that a talker who compensates completely for the altered auditory feedback would have to produce a different word. Stimuli in the second condition are designed such that a talker who does not compensate at all would hear himself producing a different word. If subjects are sensitive to producing a competing word, they should compensate more in the first condition than the second. If subjects are sensitive to hearing themselves produce a competing word, they should compensate more in the second condition than the first.

The second experiment, in Chapter 5, investigates the *phoneme avoidance hy-*

pothesis, in which top-down information from the phonological inventory influences compensation for altered auditory feedback. The intuition behind this experiment is that a subject might show incomplete compensation for an F1 feedback alteration to 'head' because there are a number of competing phonemes in the vicinity of /ɛ/, and there is a high probability of hitting one of them if one compensates too much. Experiment 2 tests the phoneme avoidance hypothesis by comparing compensation for vowels in sparser and denser regions of vowel space. The expectation is that vowels in sparse regions of vowel space (e.g. /u/) will show more compensation than dense regions of vowel space (e.g., /ɛ/).

This experiment and all previous compensation experiments have found substantial intersubject variation; some subjects compensate by directly opposing the feedback shift, while some subjects follow the shift, and some do nothing at all. The final analysis investigates the hypothesis that an individual's performance is driven by the structure of his individual vowel space. Every speaker has regions of vowel space that he visits more often and regions that he occupies less often. The analysis investigates the hypothesis that compensation is more complete in "hot spots" in vowel space that a subject occupies frequently.

These experiments help to clarify the interaction between top-down and bottom-up information in speech monitoring, providing information relevant to psycholinguistic models of speech processing. The result is also relevant to models of speech motor control, which do not currently include top-down linguistic information in feedback-based adjustment of articulatory plans.

Chapter 4

Experiment 1: Word avoidance in compensation for auditory feedback shift

Compensation in Preliminary Experiment 2 was incomplete: that is, for a 250Hz shift in F1 feedback, subjects compensated by only 120Hz. Assuming that subjects have been given sufficient time to adjust to the feedback shift, the incompleteness of compensation deserves an explanation. One place to look for the "missing" compensation is in the physiology of the articulators or the auditory system, which make it more natural to produce or hear vowels in particular ways. Without precise articulatory and neurological information, it is difficult to isolate these effects. A more accessible approach is to look for top-down effects from talkers' lexicons and phonological systems on compensation for altered auditory feedback. The current chapter explores the influence of a talker's lexicon on compensation for altered auditory feedback.

The following experiment does so by divorcing somatosensory feedback from auditory feedback for nonwords with word or nonword neighbors. A difference in compensation based on the lexical status of phonological neighbors requires that the contribution of lexical information to compensation is greater than the contribution of phonological neighbors to compensation. It also requires that target regions are the same for all stimulus words, and that the incoming acoustic signal is processed the same way for all stimulus words.

The results of this experiment can inform speech motor control models by specifying what levels of language are accessible to the articulatory adjustment process. A difference in compensation between word and nonword environments would be evidence supporting an influence of the lexicon on low-level speech motor control processes. In addition, this design makes it possible to evaluate where in the speech motor control system the lexicon is accessed. Figure 4.1 shows a schematic of the speech motor control system along with two likely locations where the lexicon might be accessed. One possibility is that lexical information is accessed early and incor-

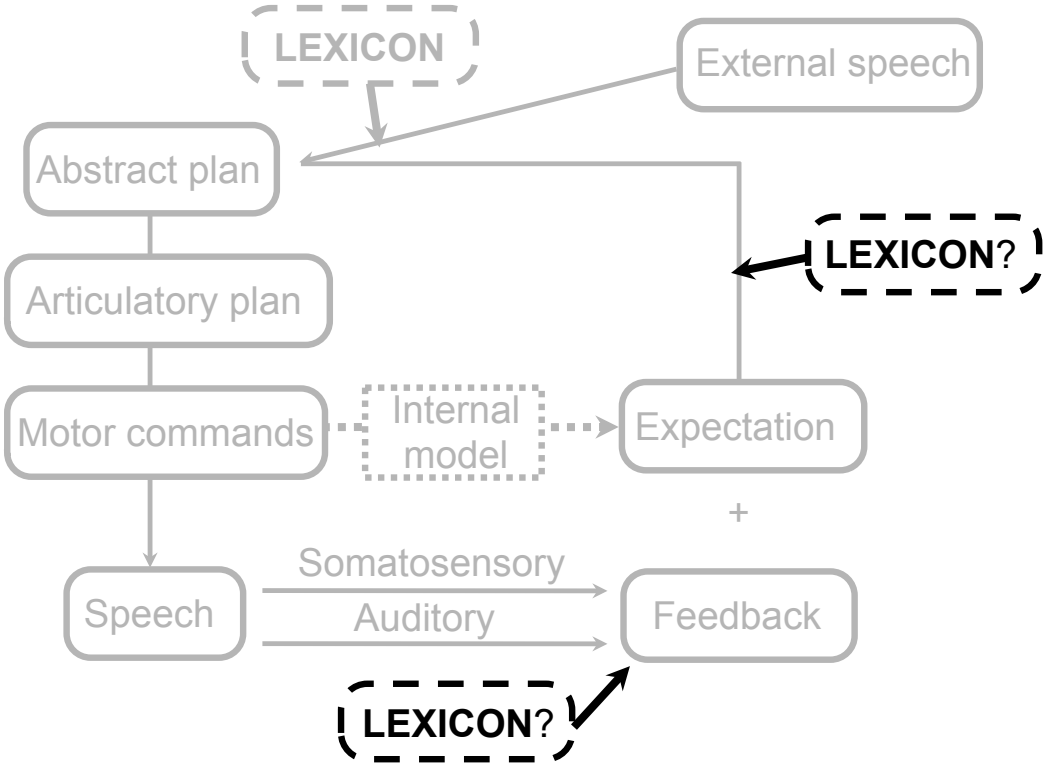


Figure 4.1: Two points in the speech motor control system where the lexicon might be accessed.

porated into a talker's auditory expectation. The Ganong effect (Ganong, 1980), which illustrates that talkers are more likely to hear words than nonwords when an ambiguous signal is presented to the ear, supports this view. Alternatively, lexical information might be accessed as compensatory articulatory plans are constructed, *after* perception and comparison of observed to expected feedback. This late-lexical-access view predicts a production-side effect in which talkers are more keen to avoid producing words than nonwords during online compensation.

Materials and methods

In order to maximally control for effects of phonology and perception, stimuli were chosen to be as similar as possible while still having different lexical neighbors. The two nonword minimal pairs that were chosen, 'deg' and 'teg', differ only in the voicing of the initial consonant. Because 'deg' and 'teg' are a minimal pair, it is minimally likely that their vowel target regions will differ or that their vowels will be perceived differently.

In this experiment, as subjects produce the stimulus words, their auditory feed-

back is shifted such that producing an /ɛ/ results in an /ɪ/ percept. Subjects producing 'deg' hear themselves saying 'dig', and subjects producing 'teg' hear themselves saying 'tig'. In both cases, somatosensory feedback is not altered, and they continue to feel themselves saying 'deg' or 'teg'. If subjects try to avoid *hearing* themselves saying a neighboring word, they should compensate by opposing the formant shift more in 'deg' than in 'teg'. Subjects can oppose this formant shift by producing the vowel /æ/; when /æ/ is altered by the FAD, it sounds like /ɛ/. 'deg' subjects can compensate completely by producing the nonword 'dag', and 'teg' subjects can compensate completely by producing the word 'tag'. If subjects want to avoid *feeling* themselves producing a neighboring word, they again should oppose the formant shift more in 'deg' than in 'teg'. This setup is illustrated in Table 4.1 and in Figure 4.2.

There are therefore two ways of seeing the same lexical effect: subjects could avoid receiving word auditory feedback from 'dig', or avoid receiving word somatosensory feedback from 'tag'. The lexical interference hypothesis predicts that compensation for altered auditory feedback should be greater for altered 'deg' than for altered 'teg'. A lack of difference in compensation between these two nonwords would be consistent with a observed-expected comparison system that does not incorporate lexical information.

Heard (w/ no compensation)	Target word	Produced (complete compensation)	Avoid hearing word	Avoid feeling word
/tɪg/	/tɛg/	"tag"	compensate less	compensate less
"dig"	/dɛg/	/dæg/	compensate more	compensate more

Table 4.1: Summary of experimental manipulation. At the beginning of the experiment, subjects say and hear themselves say 'deg' or 'teg'. If subjects do not compensate at all, they hear themselves saying /tɪg/ or 'dig' at the maximum shift. If subjects compensate completely, they produce (and receive somatosensory feedback reflecting) 'tag' or /dæg/.

Methods

Subjects (n=7, all males) sat in a sound booth. They wore a headset wired so that microphone input is fed through a computer and back into its earphones in real time. During three sessions on three separate days, they produced isolated monosyllabic words under the following conditions.

1. Control. Subjects produced {b,c,d}Vd words presented visually on the screen

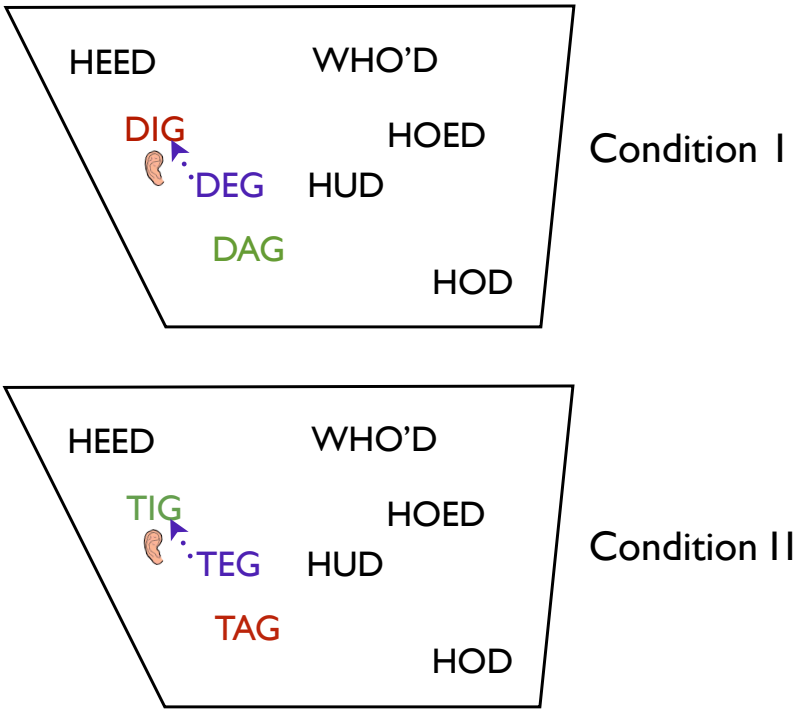


Figure 4.2: Illustration of the two experimental conditions. In Condition 1, the subject produced the nonword 'deg' as it sounded increasingly like the word 'dig'. In Condition 2, the subject produced the nonword 'teg' as it sounded more and more like the nonword /tɪg/.

(e.g., bed, bad, dad, cod). Real time feedback was played back through the earphones but was not altered.

2. DEG condition. The nonword 'deg' appeared as a prompt on a computer screen a total of 200 times, once every 2 seconds. Subjects were instructed to produce the word when it appeared.

3. TEG condition. The nonword 'teg' appeared as a prompt on a computer screen a total of 200 times, once every 2 seconds. Subjects were instructed to produce the word when it appeared.

All subjects completed the control condition first because the subject's vowel space determined the size of the formant manipulation in the other two conditions. The order of the DEG and TEG conditions was determined at random.

Formants in auditory feedback were shifted gradually in the DEG and TEG conditions from no shift up to the maximum shift. The maximum shift was equal to the distance between the centroid of the individual subject's /ɛ/ and the centroid of his /ɪ/. On average, this was a shift of about 150 Hz in F1 and 150 Hz in F2. After an initial period of unaltered feedback, auditory F1 and F2 feedback was shifted in even increments up to the maximum shift, where it was held for 90 trials. Feedback returned to normal during the last 40 trials.

Phase	**# Trials**	**Formant shift**
1	15	No shift
2	65	F1 and F2 are shifted in even increments up to the maximum shift.
3	90	Maximum F1 and F2 shift
4	40	No shift

The computer recorded both what the subject produced and what the subject heard at each trial. Recording both input and output made it possible to check that the formants were being shifted properly. Formants at each trial were measured with a script written for Entropic's ESPS utilities, as described in Chapter 3.

Results

Subjects opposed the formant shift. Figures 4.3 and 4.4 help visualize the formants that participants produced and heard over the course of the experiment. Formants in these plots were normalized by subtracting the median of each subject's first 15 (non-shift) trials from all of his subsequent trials. For each subject, the first 15 trials had normalized formant values near 0, and subsequent trials showed how much the subject deviated from baseline in response to the formant shift. Each point on the graph shows the average of these normalized values across subjects. F1 and F2 are shown in separate graphs.

The black dots in each plot show the normalized formants that were produced, on average, at each trial. The blue dots in each plot show the normalized formants that were heard, on average, at each trial. If formants were shifted properly, the sum of each black dot's position and the formant shift should be the position of the corresponding blue dot. This is usually true, but some heard formants are slightly different from the ideal values due to occasional formant mismeasurement. Because mismeasurements are relatively rare, they are not treated differently from accurate measurements.

A look at Figures 4.3 and 4.4 shows, as expected, that for the first 15 trials with no formant shift, heard and produced formants both hover around 0. From trials 16 to 80, the heard F1 decreased, and the heard F2 increased, to /ɪ/'s formants. To oppose the shifted feedback, subjects began to produce /ɛ/ with a higher F1 and a lower F2, a more æ-like vowel.

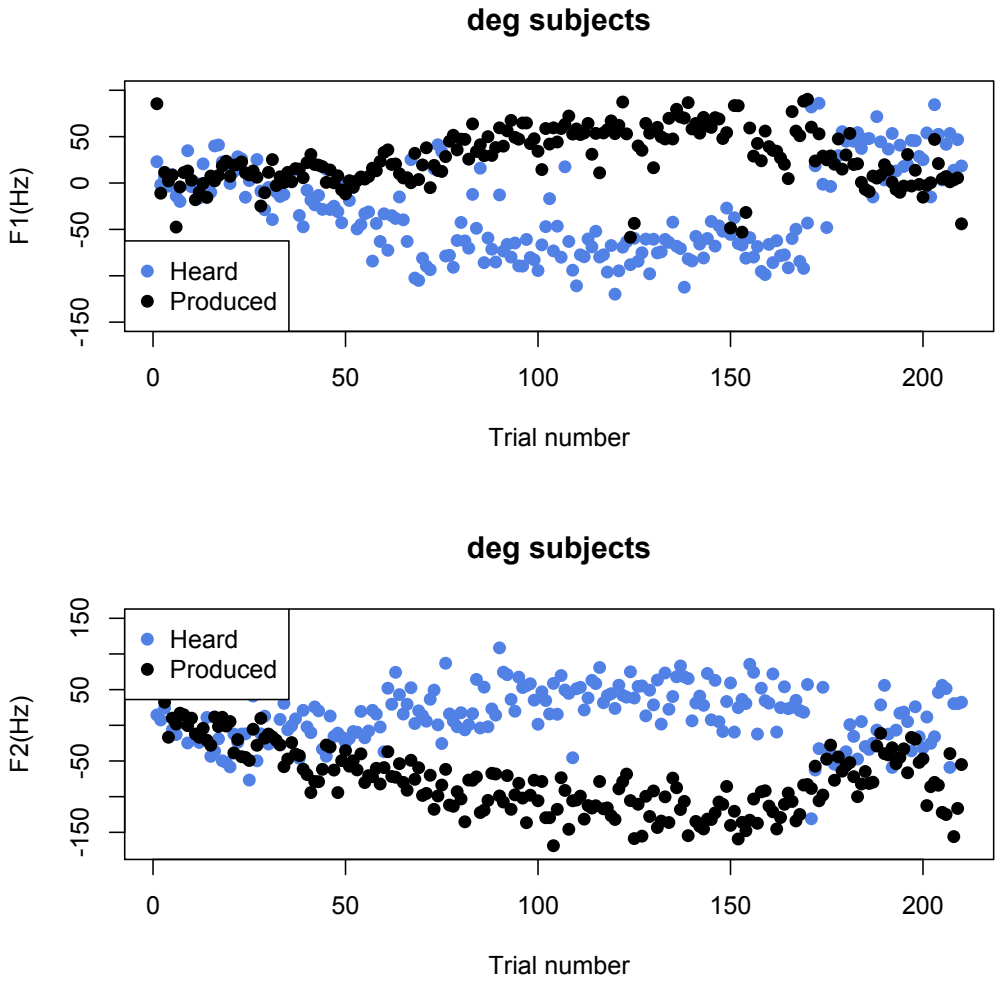


Figure 4.3: /ɛ/ formant change over the course of the 'deg' condition of the experiment. All stimuli were 'deg'. The top graph shows F1 at each trial, and the bottom graph shows F2 at each trial. Formants produced by the subjects are shown in black dots and formants heard by the subjects are shown in blue dots. Each black dot - blue dot pair represents the average of what all subjects said and heard during one trial. Each subject's measurements were scaled by subtracting the median baseline formants. Notice that subjects compensate by opposing the formant shift in both F1 and F2.

When F1 and F2 are observed separately, it is not clear whether talkers compensate more for 'deg' than for 'teg'. In the 'deg' condition, subjects change their production, on average, by 47 ± 28 Hz F1 and -113 ± 26 Hz F2, while in the 'teg'

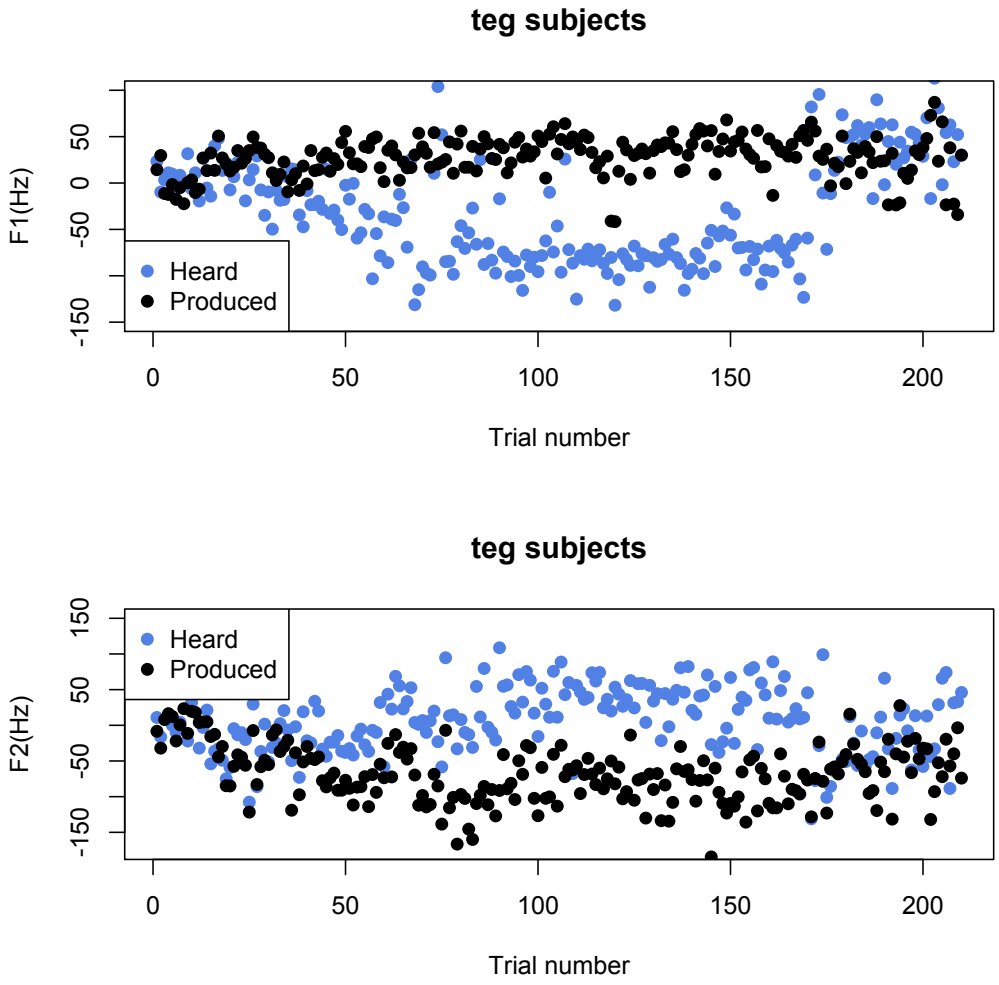


Figure 4.4: /ɛ/ formant change over the course of the 'teg' condition of the experiment. All stimuli were 'teg'. As above, the top graph shows F1, and the bottom graph shows F2 at each trial; blue dots indicate the formants that the subject heard, and black dots indicate the formants that the subjects produced. All formant measurements were scaled by subtracting the median of the first 15 (no-shift) trials.

condition, subjects change their production by about 34 ± 19Hz F1 and -85 ± 31 Hz F2. A t-test comparing the mean formant deviation during the last 40 maximum shift trials in the 'deg' condition to the corresponding 'teg' trials shows that this difference is significant for both F1 and F2 across subjects. However, this significant result might have been driven by anomalous compensation in one or two subjects; there is substantial intersubject variation. In this situation, a better method of deter-

Formant	2.5% Estimate	Mean Estimate	97.5% Estimate	$Pr(>\mid t \mid)$
F1	-14.6	-5.8	3.4	0.21
F2	**1.3**	**13.8**	**26.1**	**0.028**

Table 4.2: Summary of mixed-effect linear regression models comparing the last 40 'deg' trials with maximum shift to the corresponding 'teg' trials for each of the 5 subjects who participated in both the 'deg' and 'teg' conditions. There is no effect of condition in F1, but a significant effect of condition in F2.

mining whether there is an effect of experiment type is to run a mixed-effects linear regression. The model predicts the F2 production change at the last 40 trials with maximum shift from baseline from experiment type (deg or teg). The experimental condition is a fixed effect. To account for the fact that some subjects compensated more than others, the model includes a random effect of subject, which has two advantages: although it allows each subject to have an idiosyncratic F2 production change, it also uses the pooled data across subjects to learn about the overall production change across subjects. The model was implemented as in Chapter 3. The model was evaluated using the `lme4` package in R, and confidence intervals on the coefficients (and corresponding significance values) were calculated with MCMC sampling, using the `languageR` package in R.

The mixed-effects analysis found *no* difference between compensation in 'deg' and 'teg' conditions for F1 ($\Pr(>\mid t \mid) > 0.2$), but a possibly significant difference between compensation in 'deg' and 'teg' conditions for F2 data ($\Pr(>\mid t \mid) = 0.03$). Since this difference is in the expected direction, compensation for 'teg' may be less than compensation for 'deg', but only in F2. Because avoiding /æ/ requires producing a smaller F2, and subjects produced a higher F2 for 'teg' than they did for 'deg', their 'deg' F2 was closer to /æ/ than was their 'teg' F2. Details of the analysis are shown in Table 4.2.

Because the difference between 'deg' and 'teg' is only significant in F2, it is prudent to call the difference in compensation between 'deg' and 'teg' a weak effect. One interpretation of this result is that speakers *can* make use of lexical information in compensation for altered auditory feedback. The next section describes an experiment testing the plausibility of this interpretation. Based on the current results, it appears that speakers may move their /ɛ/ F2 further toward the nonword /dæg/ when /dɛg/ is heard as the word 'dig' than they move their /ɛ/ toward the word 'tag' when they hear /tɛg/ as the nonword /tɪg/.

The mixed-effects analysis makes one wonder whether speakers are compensating for F1 and F2 shifts independently. Vowels are a joint function of (at least) F1 and F2, and there are good anatomical reasons to believe subjects would not compensate in one formant without compensating in the other. A closer examination of Figures 4.3 and 4.4 shows that talkers change their F2 production much more than their F1

production in both conditions. In the 'teg' condition, the average maximum F1 shift that subjects hear is -75 ± 26 Hz, and they compensate only 34 Hz. On the other hand, subjects compensate -85 Hz for a 33±34 Hz maximum F2 shift. Similarly, in the 'deg' condition, subjects hear a maximum F1 shift of -83 ± 16 Hz and compensate only 47 Hz, but hear an average maximum F2 shift of 42 ± 35 Hz and compensate -113 Hz. Surprisingly, subjects are *under*compensating in F1, but *over*compensating in F2. This pattern suggests that subjects do not treat F1 and F2 separately. If they simply perceive F2 changes better than F1 changes, or can produce F2 differences more accurately than F1 differences, we might expect that F1 compensation would be smaller and more variable than F2 compensation. But the standard deviation of F2 and F1 production is similar, and compensation for F2 is greater than complete. More likely, perception and/or production operates on a joint function of F2 and F1, and talkers compensate with a joint-formant production change.

The question of whether to model F1 and F2 singly or jointly is a recurring theme in all of the experiments in this dissertation. For better compatibility with previous work, and easier visualization, it is useful to analyze F1 and F2 compensation independently. But it is important to view these analyses in light of joint data from F1 and F2, and to consider analyses that operate on both formants. In particular, because subjects are able to compensate for feedback shifts in F1 with a production change involving F1 and F2, there is a wide range of equally acceptable responses to an F1 or F2 shift. To avoid saying the word 'tag' in the 'teg' condition, talkers do not need to compensate less than they did in the 'deg' condition. Rather, they need to compensate so that their production of (F1, F2) avoids a word confusable with 'tag'. To visualize where *deg* and *teg* vowels move relative to *dag* and *tag* vowels, we need a different normalization method that can show multiple vowels on the same plot. Ideally, this method would also preserve information about the relative use of F1 and F2 during compensation.

Figure 4.5 explains the measurement method innovated here to show normalized formant movement on a coordinate plane, called a *wedge plot*. Wedge plots are essentially two-dimensional boxplots showing summary data from joint F1, F2 coordinates. Wedges have two components: the distance between the two sets of vowel measurements, and the angle between two sets of measurements in an F1-F2 coordinate plane.

A wedge plot is constructed by drawing a triangle in F1-F2 space connecting (a) the mean formants produced with no shift; (b) the mean formants produced with maximum shift; and (c) the formants that, hypothetically, would have been produced had the subject perfectly opposed the formant shift. This formulation is especially appropriate for analyzing formant data because it can answer two critical questions: (1) was compensation partial or complete? (2) did talkers oppose the auditory feedback shift directly?

The side of the triangle connecting the mean formants produced with no shift to the mean formants produced with maximum shift is the subject's *compensation*

magnitude, which shows how much the subject's production changed over the course of the experiment (question 1). The angle between the actual compensation magnitude and the ideal compensation magnitude is the *compensation angle*, which quantifies how much the subject compensated for feedback shifts in one formant with production changes in another formant (question 2).

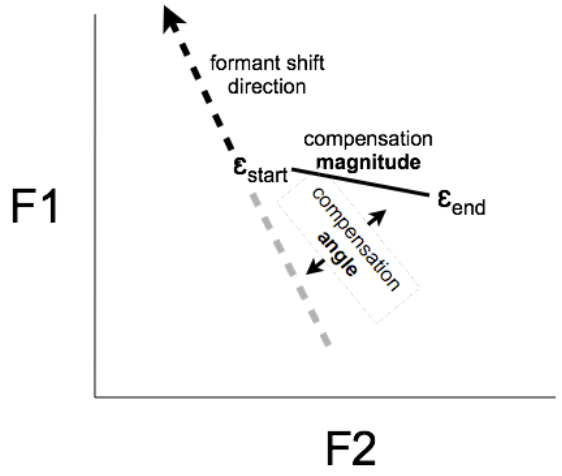


Figure 4.5: Calculations used for wedge plots. Compensation magnitude is the Euclidean distance between the formants produced during the first 15 (non-shift) trials and the formants produced during the last 40 trials of the maximum formant shift (measured in Bark). The gray dotted line shows the compensation expected if the subject opposed the formant shift directly and completely. The compensation angle is the angle between the expected end formants and the average formants actually produced during the last 40 maximum shift trials.

The wedge plots used here compare talkers' baseline vowel production, as measured by the median of the first 15 unaltered trials of the experiment, to the mean formants produced during the last 40 trials of the maximum shift condition[1]. It is perhaps useful to think of the baseline measurement as the origin of a coordinate plane, and the (distance, angle) measurement as the distance from that origin expressed in polar coordinates. The inner and outer bounds of the wedge mark the 25th and 75th magnitude measurement percentiles, and the lower and upper radii mark the 25th and the 75th percentile angles. Each wedge therefore encompasses 50% of magnitudes and 50% of angle measurements.

[1]Although extreme outliers were removed from the data, even small errors in measurement are likely to skew the mean of only 15 trials. For this reason, the median, rather than the more typical mean, was used to measure the middle of the formant distribution for the unaltered trials. Because there were 40 trials measured in the maximum shift condition, skewing from a single unusual data point was less of a concern.

A wedge plot showing summary data for the *deg* condition and the *teg* condition is shown in Figure 4.6. Color-coded diamonds show the average altered vowel that was *heard* at maximum shift in the two conditions. The plot shows a tiny difference between 'deg' and 'teg' conditions which, as in the time course data, is not present in most individuals. Individual wedge plots, shown in Figure 4.7 confirm that only subject 42 responds differently in the deg and teg conditions.

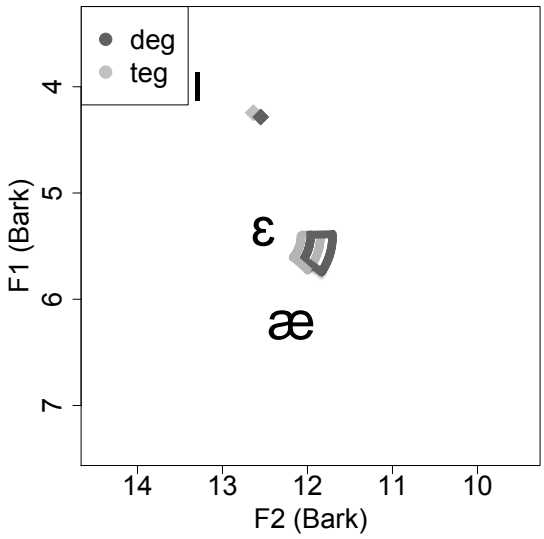


Figure 4.6: (a) Summary of results for both experimental conditions. Wedges display the middle 50% of formants produced across subjects during the last 40 trials with maximum formant shift. Formants were normalized with the magnitude-angle method described above. The darker wedge shows data from the 'deg' condition, and the lighter wedge shows data from the 'teg' condition.

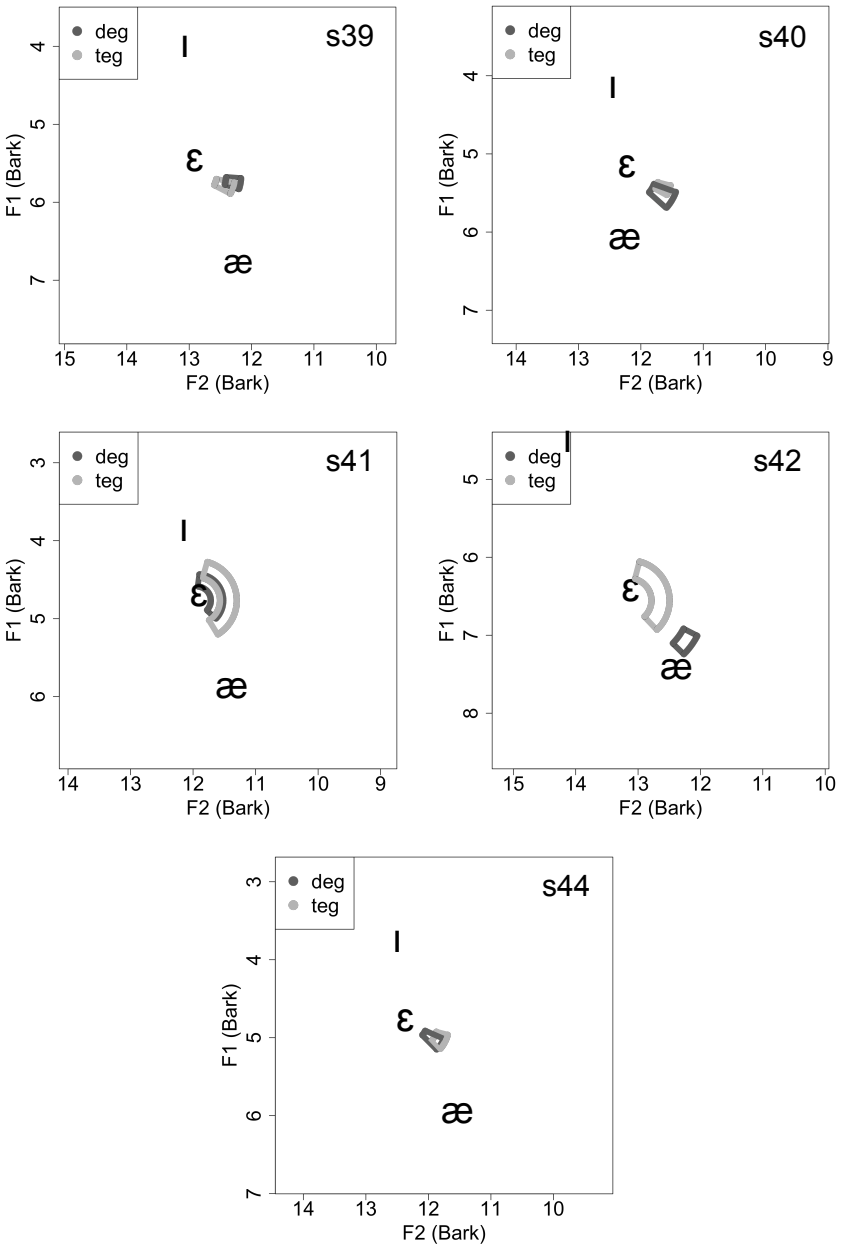


Figure 4.7: Individual results for 'deg' and 'teg' experimental conditions. Wedges display the formant difference between the first 15 unaltered trials and the last 40 trials from the maximum shift condition. The inner and outer bounds of the wedge mark the 25th and 75th magnitude measurement percentiles, and the lower and upper radii mark the 25th and the 75th percentile angles (see text for a more complete description). The darker wedge shows 'deg' data, and the lighter wedge shows 'teg' data. Only subject 42 shows a difference between conditions.

Importantly, Figure 4.6 also shows what subjects *hear*. This allows us to see whether talkers avoided saying adjacent words, avoided hearing adjacent words, both, or neither. Figure 4.6 shows that, in the aggregate, subjects hear a vowel slightly closer to /ɪ/ in the 'tig' condition and say a vowel slightly closer to /æ/ in the 'dig' condition, as predicted by the lexical avoidance hypothesis. But neither produced vowel is especially close to /ɪ/ or /æ/ because subjects compensate so much more in F2 than F1. This compensation strategy fails to support a lexical avoidance hypothesis in production or perception, but may suggest that subjects are sensitive to adjacent phonemes.

This subtle deg-teg difference is not a consequence of retaining only the middle 50% of measurements for all subjects. The difference remains when 95% of measurements from all subjects are included, as shown in Figure 4.8.

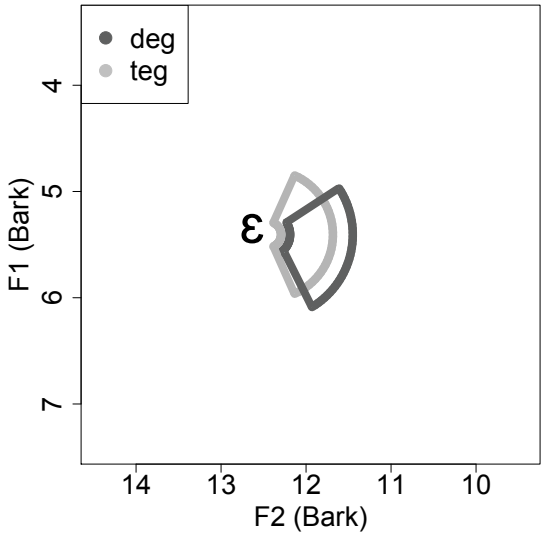


Figure 4.8: (a) Summary of results for both experimental conditions. Wedges display the formant difference between the first 15 unaltered trials and the last 40 words from the maximum shift condition. The inner and outer bounds of the wedge mark the 2.5th and 97.5th magnitude measurement percentiles, and the lower and upper radii mark the 2.5th and the 97.5th percentile angles (see text for a more complete description of magnitude and angle calculations). The darker wedge shows data from the 'deg' condition, and the lighter wedge shows data from the 'teg' condition.

There are multiple explanations for the lack of difference in compensation between the 'deg' and 'teg' conditions. One possible issue is that finding a difference in compensation between *deg* and *teg* relies on talkers treating the nonwords 'teg' and 'deg' as if they were words with lexical neighbors; it is not possible for surrounding lexical items to interfere with compensation if the lexicon is not activated. Because the

experiment requires subjects to repeat stimuli 200 times, even word stimuli probably lose their meaning for subjects about halfway through the experiment. If nonwords are treated as mere syllables, a potentially real difference in compensation between the two conditions will be invisible.

Variation is also increased by asking subjects to participate in the two conditions on separate days. Subjects did seem slightly more likely to compensate on the third day of the experiment than the second day, and due to the small number of subjects, small effects of day or time of day might be magnified. There might also be something about the specific word pair that was tested such that talkers are particularly eager to avoid hearing *dig* or saying *tag*.

To get a better understanding of the observed difference in compensation, four additional subjects were run with an experimental design that measured compensation for 'deg' and 'teg' in the same experiment and encouraged use of the lexicon.

Communicative deg/teg task

In the communicative deg/teg task, two novel objects were placed on a table and labeled either "This is a TEG" or "This is a DEG". The experimenter sat in front of the objects. Subjects were seated next to the experimenter and in front of a computer screen. Subjects were informed that they would see the word *teg* or *deg* appear on the screen, and that they were to tell the experimenter to pick up the object whose name they see. The experimenter would then pick up the object, and the subject would mark the experimenter as either correct or incorrect using a score sheet. The experimenter was cooperative and generally missed 0-2 objects per 200 word session.

Formant manipulation proceeded as in the previous deg/teg experiment, except that /ɛ/ formants were shifted toward /æ/ rather than /ɪ/. That is, subjects initially heard themselves saying 'deg' or 'teg' and eventually heard themselves saying 'dag' or 'tag'. Subjects would compensate by producing a vowel closer to /ɪ/. If there is an auditory word avoidance effect, subjects should be more inclined to change their formant production for 'teg' (to avoid hearing 'tag') than they would be for 'deg' (because they should not avoid hearing /dæg/). Table 4.3 summarizes the manipulation.

Communicative deg/teg results

To observe formant differences based on word, 'deg' and 'teg' trials were extracted from the experiment and analyzed separately. Time series data for 'deg' were normalized by subtracting the median of the initial unaltered 'deg' trials for each subject, and time series data for 'teg' were normalized by subtracting the median of the initial unaltered 'teg' trials for each subject. The formants heard and produced across subjects are shown in Figures 4.9 and 4.10. Because 'deg' and 'teg' stimuli were ordered randomly for each subject, the x-axis in these figures no longer shows

Heard (w/ no compensation)	Target word	Produced (complete compensation)	Avoid hearing word	Avoid feeling word
"tag"	/tɛg/	tɪg	compensate more	compensate more
/dæg/	/dɛg/	"dig"	compensate less	compensate less

Table 4.3: Summary of experimental manipulation. At the beginning of the experiment, subjects will say and hear themselves say 'deg' or 'teg'. If subjects do not compensate at all, they will hear themselves saying 'tag' or /dæg/ at the maximum shift. If subjects compensate completely, they will produce (and receive somatosensory feedback reflecting) /tɪg/ or 'dig'.

trial number; it shows the instance of the stimulus word. For example, the 3rd black dot in the 'deg' plot shows the average F1 or F2 that was produced the third time that 'deg' appeared on the computer screen. For different subjects, this would have occurred at a different trial number. For this reason, aggregate measurements need to be interpreted with some caution.

As in the noncommunicative version of this experiment, it is unclear from the raw timeseries data whether subjects compensated differently for *deg* than for *teg*. Compensation for the two conditions averaged across subjects are shown below. Figure 4.9 shows the formants that talkers produced and heard during 'deg' trials. In F1, they compensate -24±14Hz for a heard shift of 86±17Hz, and in F2, they compensate 36±25Hz for a maximum shift of -27±42Hz. Figure 4.10 shows what subjects produced and heard during the 'teg' stimulus trials of this experiment. For these trials, subjects compensated only -13±13Hz for an average F1 shift of 93±10Hz, and compensated -1±33Hz for an F2 shift of -63±25Hz. The smaller compensation for the 'teg' condition suggests that subjects were not uncomfortable hearing 'tag' for 'teg'.

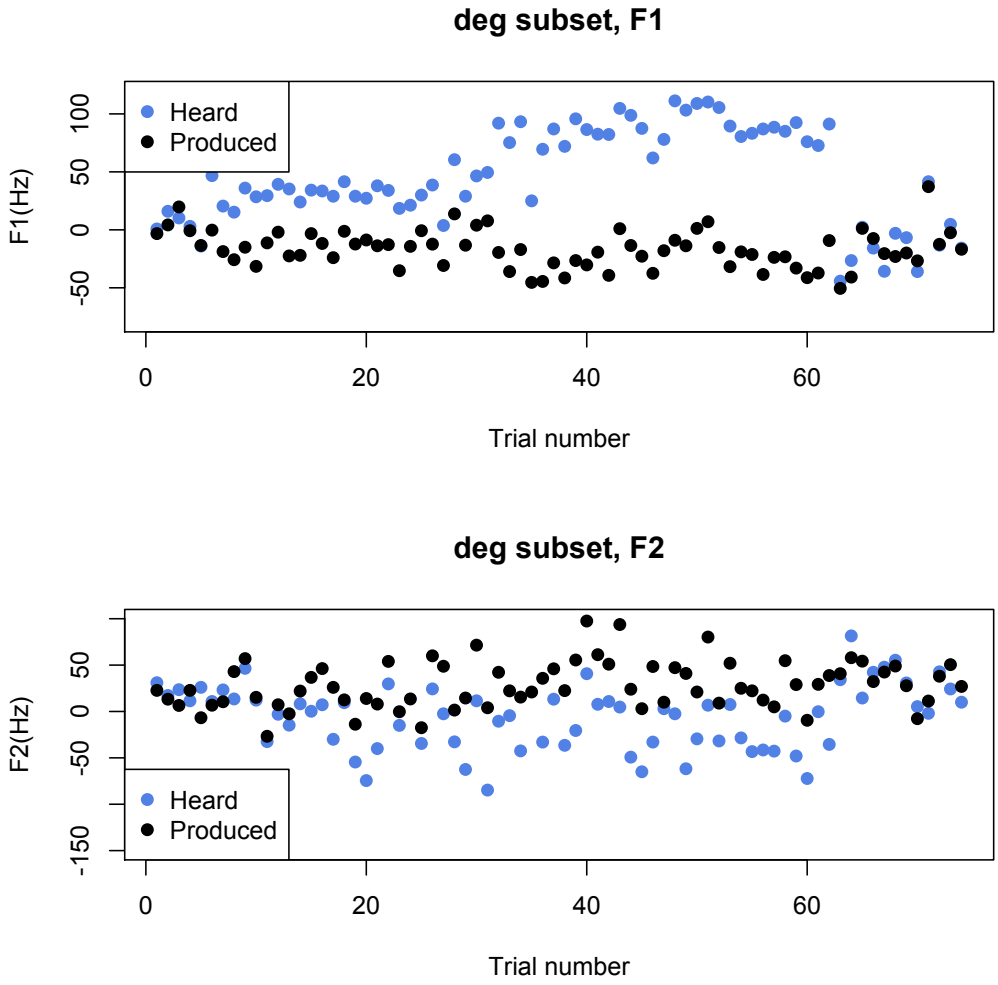


Figure 4.9: /ɛ/ formant change during the the Communicative deg/teg task. Averages for 'deg' stimuli are shown. As above, the top graph shows F1, and the bottom graph shows F2 at each trial; blue dots indicate the formants that the subjects heard, on average, and black dots indicate the formants that the subjects produced, on average. All formant measurements were scaled by subtracting the median of the first 15 (no-shift) trials.

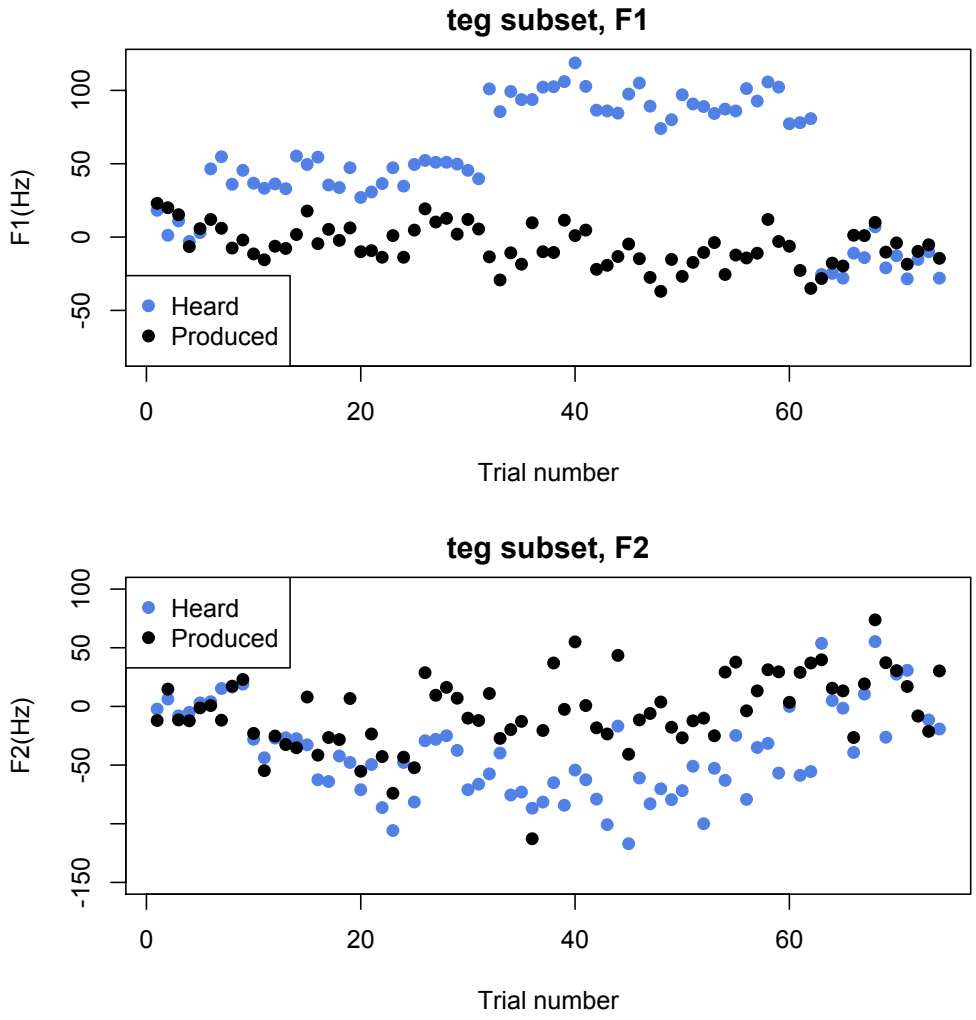


Figure 4.10: /ɛ/ formant change during the the Communicative deg/teg task. Averages for 'teg' stimuli are shown. As above, the top graph shows F1, and the bottom graph shows F2 at each trial; blue dots indicate the formants that the subjects heard, and black dots indicate the formants that the subjects produced. All formant measurements were scaled by subtracting the median of the first 15 (no-shift) trials.

The wedge plot in Figure 4.11 shows that for these talkers, /æ/ is quite close to /ɛ/. Though there is no difference in compensation magnitude between productions of 'deg' and productions of 'teg', there is some difference in the range of compensation angles; there is a larger variety of angles for 'deg' than for 'teg'.

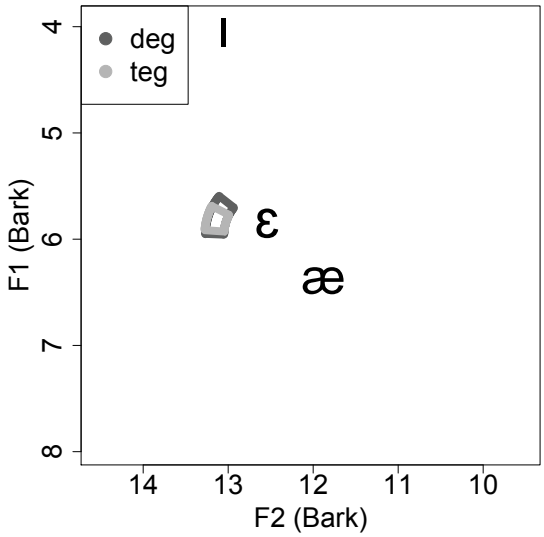


Figure 4.11: /ɛ/ formant change during the the Communicative deg/teg task. The average interquartile range (25% - 75%) for 'deg' and 'teg' stimuli are shown. Across subjects, there is no difference in compensation magnitude for the two stimuli, and a small difference in the compensation angles for the two stimuli. Neither is at all confusable with adjacent vowel /ɪ/.

Although there is no difference between 'deg' and 'teg' formants across subjects, individual wedge plots show that subjects 49 and 52 have significantly different formants for the two words. Because any given instance of 'deg' is likely to come right before or after an instance of 'teg', and the subject's articulators are primed to produce similar formants in adjacent trials, it is surprising to find that any subject produces the two words differently.

However, the individual graphs shown in Figure 4.12 fail to support the lexical interference hypothesis: the four subjects in this experiment respond to altered auditory feedback in four different ways. No subject compensates by producing a vowel remotely confusable with /ɪ/. Moreover, it is not clear that any subject compensates so as to produce a more /ɪ/-like vowel for 'teg' than for 'deg'. Subject 52 compensates more for 'teg' than 'deg', but the formant shift he heard was unusually large because /ɛ/ and /æ/ are particularly far apart in his vowel space. Subject 50 seemed to respond similarly to altered feedback in the two words. Subject 49 compensated more for 'deg' than 'teg', *contrary* to the lexical avoidance hypothesis. It is not clear

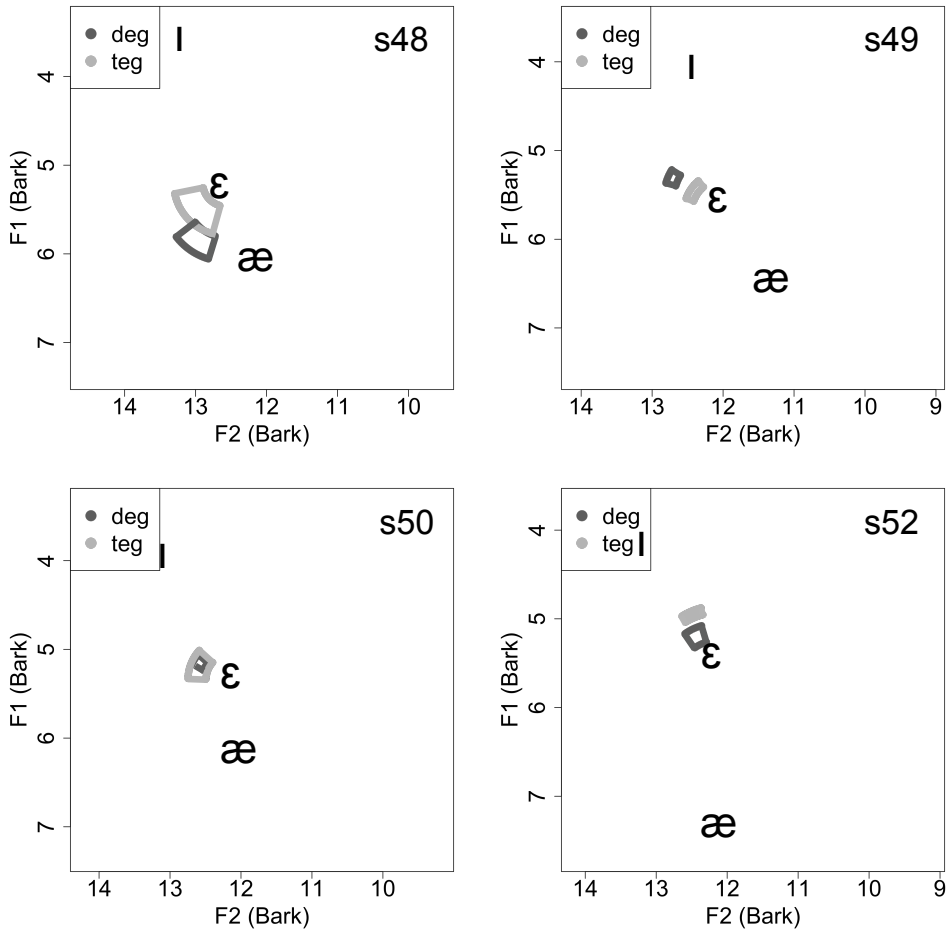


Figure 4.12: Individual results for Communicative 'deg' / 'teg' task. Wedges display the formant difference between the first 15 unaltered trials and the last 25 words from the maximum shift condition. Wedges mark the 25th and 75th magnitude and angle percentiles (see text for a more complete description of calculations). The darker wedge shows data from the 'deg' condition, and the lighter wedge shows data from the 'teg' condition.

how to characterize subject 48's performance; he neither followed nor opposed the formant shift. With such a broad spectrum of responses to the feedback shift, it is very unlikely that subjects compensated more for 'teg' than for 'deg', however, it is worth analyzing the data as a whole to characterize any cross-subject patterns that exist.

To do so, cross-subject patterns were measured with two mixed-effect linear regression analyses. As in the non-communicative version of this experiment, one measured the role of stimulus word in predicting F1 production, and the other mea-

Formant	2.5% Estimate	Mean Estimate	97.5% Estimate	$Pr(>\mid t \mid)$
F1	**10.9**	**18.3**	**25.7**	< 0.0001
F2	**-58.9**	**-43.8**	**-28.5**	< 0.0001

Table 4.4: Summary of mixed-effect linear regression models comparing the 'deg' trials with maximum shift to the 'teg' trials with maximum for all subjects in the communicative version of the deg/teg experiment, with 'deg' as baseline. Compensation for the two stimulus words are significantly different in both formants, though **not** in the expected direction.

sured the role of stimulus word in predicting F2 production. Like the t-test performed above, these two models used the stimulus word ('deg' or 'teg') as a fixed effect to predict F1 or F2 production change. Subject was included as a random effect, which effectively reduced the contribution of outliers. Data included all instances of 'deg' and 'teg' with maximum shift. The effect size and significance of the two stimulus words, with 'deg' as baseline, is shown in Table 4.4.

This test shows that the 'deg' and 'teg' stimuli are associated with different amounts of compensation. There is a very significant lexical effect ($Pr(>\mid t \mid) < 0.0001$). However, Table 4.4 shows that this effect is not in the expected direction. Recall that subjects were expected to compensate more in the direction of nonword /tɪg/ than real word 'dig', meaning that subjects should produce a *lower* F1 and a *higher* F2 in 'teg' stimuli. Notice that the coefficient of the 'teg' stimuli is positive in F1 and negative in F2. Counter to the lexical interference hypothesis, subjects are compensating more for 'deg' than for 'teg'.

To verify this result, Figure 4.13 shows the difference in compensation between the two stimuli. The histogram was created from mixed-effects model with stimulus as fixed effect and subject as random effect, but with no intercept. In this case the model estimates the mean and variance of compensation for the two stimuli. Confidence intervals for these mixed-effects models are constructed by running the pvals.fnc MCMC sampler, which generates 'deg' and 'teg' distributions from 10,000 random draws. To decide whether the 'deg' and the 'teg' groups are equivalent, the 'deg' samples were subtracted from the 'teg' samples. Figure 4.13 shows histograms of the difference between the 'deg' samples and the 'teg' samples for F1 and F2, respectively.

As explained above, the lexical interference hypothesis predicts that subjects' change in F1 production should be more negative for 'teg' than for 'deg', and subjects' change in F2 production should be more positive in 'teg' than in 'deg'.

The histograms in Figure 4.13 confirm that these data do not support the lexical interference hypothesis. They show clearly that (1) the difference between 'teg' and 'deg' is not 0; (2) the difference between 'teg' and 'deg' is positive for F1; (3) the difference between 'teg' and 'deg' is negative for F2. Again, contrary to the lexical

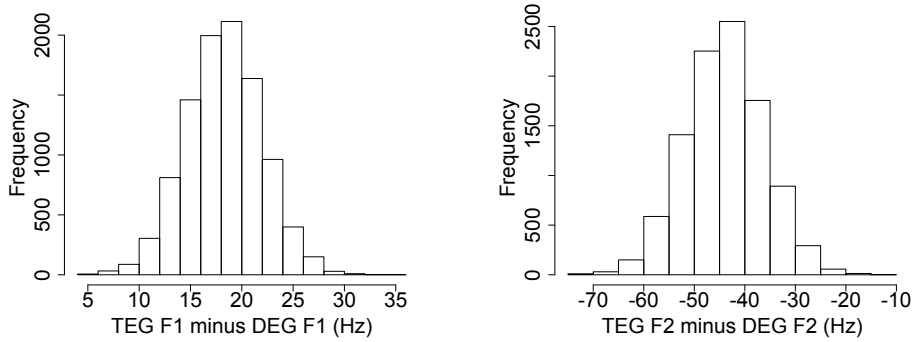


Figure 4.13: (Left) Histogram of 10,000 'teg' F1 samples minus 10,000 'deg' F1 samples from MCMC. (Right) Histogram of 10,000 'teg' F2 samples minus 10,000 'deg' F2 samples from MCMC. Neither histogram supports the lexical interference hypothesis.

interference hypothesis, subjects compensate *more* for 'deg' than for 'teg' in F1 and in F2. It is not clear why these subjects compensated more in the nonword environment than in the word environment.

Discussion

This pair of experiments set out to determine whether the lexicon is integrated into the speech motor control system, and if so, at what stage of processing. The first deg/teg experiment found that subjects avoided saying 'tag' marginally more than 'dag' (or avoided hearing themselves say 'dig' marginally more than 'tig') when F1 and F2 were evaluated separately. This is best interpreted as weak support for the lexical interference hypothesis.

The communicative deg/teg task was a stronger test of the lexical interference hypothesis: in that task, subjects produced 'deg' and 'teg' in a communicative game setting, maximizing the benefit to the listener of distinguishing between the two words. In the communicative task, there was no consistent lexical interference effect. Instead, individual graphs showed substantial intersubject variation: two subject opposed the shift; one subject did not compensate; and one subject behaved unexpectedly. A mixed-effects linear regression model including subject as a random effect demonstrated that, contrary to the lexical interference hypothesis, subjects compensated more in the nonword condition than in the word condition.

Though lexical avoidance is not likely to be driving responses to the auditory feedback shift, more analysis is needed before a firm conclusion can be drawn. Not only is there a lexical effect for F2 data in the non-linguistic version of this experiment, but the wedge plots in Figures 4.7 and 4.12 show that F1 and F2 are processed jointly. An analysis based on some yet-to-be-determined function of F1 and F2 would likely

yield a more productive understanding of this data.

If the lexicon is integrated into the speech motor control system at either the pre-comparison or the pre-planning stage, it probably is integrated optionally. Otherwise, these two experiments ought to have yielded stronger lexical effects. Although the lexicon may not affect speech motor control directly, it is still obviously active at higher levels of speech planning. These experiments simply suggest that speech plans pass through the lexicon before they reach the part of the speech motor control system that deals with low-level feedback. Several predictions naturally follow from this result. If there is no lexical influence on the speech motor control system, the speech motor control system is unlikely to have a direct influence on the lexicon or lexical storage, and learning or adaptation resulting from this experiment should not be lexically specific. Specifically, compensation for altered auditory feedback should generalize to homophones, and to any other word sharing an articulatory planning unit with the altered word(s).

Experiments involving altered somatosensory feedback show that, in addition to articulatory planning units, speech mode matters. Tremblay, Houle, and Ostry (2008) stabilized speakers' heads in a device that briefly pulled their jaws to the left or right during speech. The jaw-pulling had no effect on speech acoustics. The authors found that *somatosensory* adaptation to the nonword /siæs/ does not transfer either to a larger nonword, /siæis/, containing the test nonword, or to a silently-mouthed version of the test nonword /siæs/. In other words, subjects failed to generalize the learned adaptation to new words containing the adapted syllables. If feedback were monitored on either the phoneme or on the syllable level, adaptation would be visible in other words sharing the adapted word or syllable. Not only does adaptation fail to spread to new words with some identical syllables, but adaptation also fails to spread to instances of the same word said in a new speech mode (in the Tremblay et al. case, silent mouthing). Apparently, subjects are sensitive to their environmental context (i.e., whether they are in silent or voiced speech mode) when compensating. It remains to be tested whether compensation generalizes to homophones in the same speech mode, controlling for environmental context, though chapter 6 touches on this issue.

Some caution is warranted in interpreting these results; though no clear lexical effect was found, there is also no direct evidence that a lexical effect is absent. Though the failure of the communicative task to find a lexical effect is suggestive, it remains possible that for some perceptual reason, the shifted version of /tɛg/ is not confusable with 'tag', or the compensated production of /dɛg/ is not confusable with 'dig'. If there is no danger of hearing or producing a lexical neighbor, we should not expect a difference between conditions. The best test of this possibility would be to have subjects come in on a separate day and classify the words they heard and produced.

As an additional caution, it is important to recall that this experiment tested lexical influences on compensation, but not on adaptation. It is possible that adaptation is lexically influenced, even if compensation is not. Compensation is a reaction of the

speech motor control system with fast onset and fast offset. Adaptation, a learned change in speech that remains after feedback returns to normal, may be subject to category biases that are not present in lexical contexts. Generalization experiments measure the degree to which adaptation in one vowel spreads to other words with the same vowel or to other vowels. Those who find no generalization to other words with the same vowel must conclude that there is lexical interference on learning.

With this study completed, we can narrow the list of explanations for partial compensation introduced at the beginning of the chapter. The results presented here are not consistent with a lexical interference effect. They are instead consistent with phoneme avoidance, or alternatively, articulatory habits or perceptual warping. The consistency with which subjects overcompensate in F2 but undercompensate in F1 is still in need of investigation, and the huge individual differences in strategy remain unexplained. The following experiments take up both of these issues.

Chapter 5

Experiment 2: Somatosensory and phonemic influences on compensation for shifts in auditory feedback

Experiments altering auditory feedback have shown that talkers will change the way they articulate vowels in order to preserve their acoustics (Houde & Jordan, 2002; Purcell & Munhall, 2006b; Katseff, Houde, & Johnson, 2008). Experiments altering somatosensory feedback have shown that talkers will change their motor plans for target vowels in order to preserve their articulatory configurations (Tremblay et al., 2003; Nasir & Ostry, 2006; Tremblay et al., 2008). Chapter 3 of this dissertation confirmed that that talkers will change their articulation of target vowels in the presence of altered auditory feedback, but additionally found that for large shifts, compensation is too small to preserve the acoustics of the target vowel. Drawing on this evidence, I argued that a vowel is defined by both a set of acoustic expectations, comprised of a set of formants in vowel space, and a set of somatosensory expectations, comprised of muscle length and velocity targets from active articulators accompanied by pressure targets from the passive articulators.

If vowel targets have acoustic and somatosensory components, in an altered feedback experiment there is no way to keep both types of feedback on target. Because altered feedback experiments require speakers to learn to accept a new acoustic target or a new somatosensory target – or both – one can investigate what is driving them to change their production – or fail to change their production – in both domains.

There is substantial evidence that vowels are associated with low-level expectations from the somatosensory and auditory domains, and it is possible that incompleteness and indirectness in compensation are due purely to the relative weighting of these two sorts of feedback. Alternatively, high-level expectations about a talker's language could also influence compensation for altered auditory feedback above and

beyond the contribution of low-level somatosensory and auditory feedback. Chapter 4 investigated this hypothesis by looking for lexical effects on compensation for altered auditory feedback. That experiment did not find any consistent, cross-talker effects from the lexicon.

Because psycholinguistic models of speech perception suppose that all incoming speech is routed through the phonological inventory, it is logical to continue this investigation with phonemes. According to these models, deciding how much to compensate for altered auditory feedback should depend not only on spectral and constriction differences between the expected vowel and the actual vowel, but also on whether the heard or produced vowel is an acceptable instance of the target phoneme. Vowels altered to a position within a phoneme category ought to be treated differently from vowels altered to a position outside the normal range of phonetic variation for that phoneme category.

If vowel targets have acoustic and somatosensory components, there are two sources of uncertainty defining the target range in each domain. If a vowel is defined as the articulatory positions that allow a talker to achieve a certain set of formants, then its limits are constrained by the Quantal Theory of Speech (Stevens, 1972, 1989), which states that certain vowels – particularly the point vowels /i/, /a/, and /u/ – exhibit minimal acoustic variability for relatively large changes in articulatory position. This theory predicts that vowels used in languages are stable points in acoustic space. While California English /i/, /a/, and /u/ may have a more limited range of articulation than other vowels, these vowels are not located far enough in the corners of vowel space to exhibit saturation effects (Fujimura & Kakita, 1979). The /u/ vowel is particularly far from the corner of vowel space (Hagiwara, 1997; Hall-Lew, 2009).

If a vowel is instead defined by a certain set of somatosensations received by the central nervous system from the vocal tract, then its limits are defined by the precision and accuracy of vocal tract mechanoreceptors. Through mechanoreceptors on the palate, lips, and tongue, speakers can detect palatal contact and lip rounding. Jaw and vocal fold mechanoreceptors provide additional information about the state of the vocal tract over time. This somatosensory information is available to the central nervous system and is used, in conjunction with auditory information, to form speech targets[1].

Preliminary Experiment 2, presented in Chapter 3, noted that compensatory production appeared to approach an asymptote as the amount of feedback shift increased. Recent research in this area has shown that compensation does indeed hit an asymptote for a large enough F1 shift (MacDonald et al., 2010). The explanation presented in Chapter 3 suggested that the discrepancy between observed and expected auditory feedback causes talkers to rely on the more consistent somatosensory feed-

[1]Because it is not certain whether these targets are phonemes, allophones, syllables, words, phrases, or some combination of these, it is premature to say they form phoneme categories, but this is certainly a possibility.

back. This is equivalent to an account where talkers are willing to alter their acoustic targets more than they are willing to alter their somatosensory targets. Of course, this result is based on altered feedback in a single vowel with singular acoustic and somatosensory characteristics. A natural test of this hypothesis would be to compare compensation for altered auditory feedback in vowels with a range of auditory and somatosensory attributes.

One such attribute is proximity of nearby competitor vowels. To get a sense for the impact of competitor vowels on the speech motor control system, consider a perturbation to auditory feedback in a hypothetical English speaker whose vowel in 'bed' is similar to the vowel in 'bad'. There are two situations in which acoustic (as opposed to somatosensory) phoneme targets might influence compensation for altered auditory feedback.

First, the phoneme inventory might influence compensation when processing incoming auditory feedback by way of a categorical perception effect. Consider a situation where a hypothetical speaker hears his 'ged' vowel altered to [æ]. Because the spectral difference between what he hears (gad) and what he expects to hear (ged) is small, his speech production system might attribute the mismatch to noise, which would make him less likely to compensate for the altered feedback. On the other hand, the altered [ɛ] sounds like a vowel in a different category. If that makes the altered feedback more salient, the speaker might be more likely to compensate for the altered feedback.

Second, the phoneme inventory might influence compensation when deciding how to compensate for altered auditory feedback. Consider the opposite case, in which 'ged' is altered such that the vowel that the talker *hears* is still classified as an /ɛ/, but he would have to *produce* 'gad' in order to oppose the change. Although the production change required to compensate completely is quite small, it requires crossing a category boundary: full compensation would require him to say 'gad' when he meant 'ged'. This speaker might be less inclined to compensate for this auditory feedback shift than for an equally-sized shift that he could compensate for without crossing a category boundary (for example, a shift of the same size in the neighboring nonword 'gid').

In other words, we can learn about whether a talker's phoneme inventory interacts with his speech motor control system by looking for differential responses to feedback alterations of the same size in different phonemes, and by observing the correlation between acoustic and somatosensory targets and compensation for missing those targets. Target-based expectations might affect either the range of vowels one is willing to say, or the range of vowels one is willing to hear, as depicted in Figure 5.1.

This chapter looks for a phonological influence on compensation in the same way that the previous chapter sought to find a lexical influence on compensation. As in the case of lexical influence, there are two likely locations where information from the phonological inventory might influence compensation. In one case, auditory

and somatosensory feedback might be incorporated with phonological information, changing the magnitude of the mismatch between observed and expected feedback. That is, combining the actual mismatch with one's phonological inventory will mean mismatches that change the categorization of a phoneme would be more salient than mismatches of equal size that do not cross a category boundary.

There is already a profusion of evidence that perception is routed through phonological categories. Categorical perception effects show that participants tend to classify stop tokens unambiguously, even when they fall halfway between two phonemes (Liberman, Harris, Hoffman, & Griffith, 1957; Liberman et al., 1967). Discrimination effects, in which participants are maximally able to distinguish two similar sounds along a continuum when those sounds fall into different categories, also show that talkers tend to judge what they hear through the lens of their phonological inventories (Liberman et al., 1967; Studdert-Kennedy, Liberman, Harris, & Cooper, 1970; Pisoni & Tash, 1974). It is not known how early in auditory processing this top-down knowledge impinges on auditory perception.

In addition to influencing the perceived mismatch between observed and expected feedback, the phonological inventory might influence the change in articulatory plan made to compensate for the altered auditory feedback. Rather than exactly oppose the formant shift, talkers may prefer to produce vowels that are clearly a member of a vowel category, or maximally likely to be heard as the intended vowel, even if that requires producing a vowel that does not directly or completely oppose the formant shift.

This experiment tests responses to altered auditory feedback for three different vowels: /ɛ/, /u/, and /ʌ/. Talkers' individual and group responses to alterations in these vowels will be considered with respect to nearby vowel competitors and salience of somatosensory feedback.

Viewing vowel targets as having auditory and somatosensory components leads to two straightforward predictions for how compensation will vary from vowel to vowel. The presence of acoustically similar nearby vowels increases the importance of getting vowel acoustics right. American English /ɛ/ is a good example of such a vowel; it is adjacent to three vowels, /ɪ/, /e/, and /æ/, that are so similar that they are targets of vowel mergers in some dialects (e.g., the pin/pen merger in the North American Midlands region, in which /ɪ/ and /ɛ/ are merged before nasals /m/ and /n/ (Ash, 2006; Labov, Ash, & Boberg, 2006; Thomas, 2004)). Alternatively, vowels with salient somatosensory feedback, either from an articulator that is used relatively rarely (e.g., lip rounding in English), or from mechanoreceptors that give straightforward information about the location of articulators and constrictions, may have targets that emphasize their somatosensory components. American English /i/ and /u/ are good candidates for vowels with salient somatosensory feedback. According to an EPG study of English vowels, /i/ has substantial palatal contact (Stone & Lundberg, 1996), which provides reliable feedback about constriction location in this vowel. The same study shows that /u/ has some palatal contact, meaning that

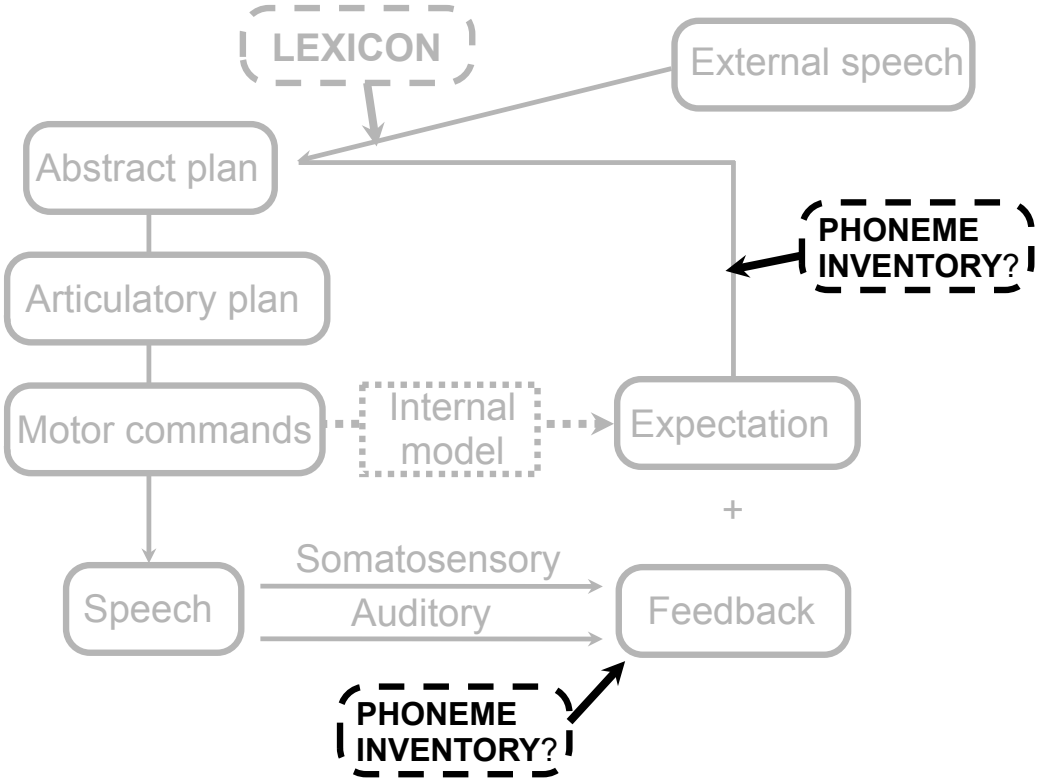


Figure 5.1: Two points in the speech motor control system where the phoneme inventory might be accessed. Contribution from the phoneme inventory to feedback makes talkers especially wary of *hearing* an out-of-category vowel. Contributions of the phoneme inventory to the adjustment process would make talkers especially wary of *saying* an out-of-category vowel.

palatal contact is somewhat informative in this vowel. More importantly, /u/ exhibits substantial lip rounding, more than any other vowel in English except, perhaps, /o/. Because few other vowels have so much lip rounding, speakers might be especially sensitive to the lip rounding in /u/. There is some controversy as to how proprioceptive articulator information is transmitted, but it appears that tongue position information is far less reliable than palatal contact information. Tongue position information must be transmitted to the central nervous system by way of muscle spindles, which simply do not exist in all parts of the tongue[2](Kuehn, Templeton, & Maynard, 1990). By contrast, palatal contact is demonstrably used in speech motor control, as shown by studies showing articulation changes in response to changes in the thickness of an artificial palate (Honda & Fujino, 2002). Thus vowels with little

[2]In particular, no spindles are present in levator veli palatini, palatopharyngeus, musculus uvulae, salpingopharyngeus, or the superior pharyngeal constrictor.

palatal contact like /ʌ/ must rely on less straightforward somatosensory feedback from muscle spindles, while vowels like /i/ can utilize more direct tongue position evidence from mechanoreceptors on the palate.

Based on the auditory component of vowel targets, one would expect phoneme boundaries to be less flexible in crowded than in sparse areas of vowel space because subjects are expected to avoid feeling themselves produce neighboring vowels[3]. A vowel like /ɛ/ would therefore be expected to show poor compensation, and a vowel like /i/ or /u/ would be expected to show more complete compensation. Based on the somatosensory portion of vowel targets, one would expect compensation to be less complete for vowels with reliable or salient somatosensory feedback, such as /i/ or /u/, and more complete in vowels with less salient somatosensory feedback, such as /ʌ/ or /ə/, because talkers might have a clear memory for how the articulators are situated during vowels with consistent feedback and might be more averse to producing these vowels with different articulations.

There is some work on the amount of variation in vowel production in citation form words (Perkell & Nelson, 1985; Beckman et al., 1995). These studies find that /i/ has substantial variation in constriction location, with slightly less for /u/, and less still for /a/ and /ɚ/. In addition, somatosensory feedback might be more salient for some vowels than others. Other things being equal, a vowel that requires lip rounding might have more salient somatosensory feedback than its unrounded counterpart. Predictions for a vowel like schwa are more complicated. Browman and Goldstein (1992) recorded articulatory movements from speakers producing vowels in a [pV$_1$pəpV$_2$pə] context and found that, though the formants in schwa varied more than formants in any other vowel, the formant range did not cover the entire vowel space, and the articulatory position of schwa could not be predicted from the previous and following consonants alone. In the context of this study, these results are consistent with an articulatory target with low salience. Speakers have an idea for where their articulators ought to be during schwa and some idea of what schwa ought to sound like, but they are not very sensitive to small deviations in acoustics or articulation. Further support for this position comes from several earlier studies (Kuehn & Moll, 1972; Fowler & Turvey, 1981; Flemming & Johnson, 2007) showing that word-medial schwa has more acoustic variation in F2 than in F1. This follows from the relative salience of feedback from jaw height relative to feedback from tongue fronting. Data from acoustic variability in English vowels suggests that a small change in /i/ articulation is noticeable, but a small change in /ə/ articulation is not. In vowels where feedback about the position of F1 comes primarily from jaw height, the salience of jaw height ought to be similar across vowels. Differences arise only when a vowel's target position requires articulators to touch each other. In this analysis,

[3]An opposing argument is also possible. If the altered vowel that one hears may either be a poor examplar of the intended vowel or a good exemplar of a competing vowel, the good exemplar of a competing vowel may be more salient. In this case, vowels in dense regions would be associated with greater compensation. This hypothesis will be considered as well.

Vowel	Lip Rounding	Palatal contact
ʌ	No	Minimal
ɛ	No	Moderate
u	Yes	Moderate

Table 5.1: Proxy for salience of somatosensory feedback in three English vowels.

somatosensory feedback will be considered *salient* in vowels involving structures that come into contact with each other (Perkell, 2007), and *weak* in vowels relying primarily on feedback from a single articulator. As a proxy for the salience of articulation of English vowels, degree of palatal contact and lip rounding is used. Information on palatal contact is taken from an EPG study of English vowels by Stone and Lundberg (1996).

Table 5.1 summarizes somatosensory feedback from the vowels /ʌ/, /ɛ/, and /u/, as measured by palatal contact and lip rounding. /ʌ/ has little palatal contact and no lip rounding, /ɛ/ has some palatal contact but not lip rounding, and /u/ has both palatal contact and lip rounding. Measured this way, multiple cues accompany a correctly articulated /u/, and different sorts of cues that are plausibly less salient accompany vowels like /ʌ/. If somatosensory feedback is more important to a vowel target when it is salient, then it would interfere most with compensation in vowels like /u/, and cause little interference in vowels like /ʌ/.

To determine the number of vowels acoustically adjacent to /ɛ/, /ʌ/, and /u/, vowel spaces were averaged across all subjects in this experiment. A circle with a diameter of 2 Bark was drawn around each vowel. There are three vowels within a 2 Bark radius of /ɛ/; five vowels within a 2 Bark radius of /ʌ/, and 1 vowel within a 2 Bark radius of /u/. The average California vowel space for speakers in this experiment is shown in Figure 5.2.

Acoustic properties of vowels also have the potential to cause interference in compensation. In auditory feedback, salience is of less concern because the most important formant frequencies in vowels occupy a small portion of the range of human hearing and are unlikely to exhibit large differences in salience. However, auditory neighbors are likely to matter. If a talker's task is to say 'head', it is logical that they might want to avoid saying 'hid' or 'had'. Although the results of Chapter 4 showed that talkers do not avoid neighboring lexical items, they also suggested that they may avoid neighboring phonemes. In other words, talkers may be just as likely to avoid saying 'gid' or 'gad' when they intend 'ged' as they are to avoid saying 'hid' or 'had' when they say 'head'. Certainly vowel formants in high density lexical neighborhoods differ from vowel formants in low density lexical neighborhoods, implying that speakers take stock of some aspects of nearby vowel space before talking (Wright, 1997, 2004; Munson & Solomon, 2004). The presence of articulatory neighbors is less important to somatosensory feedback because, at least according to Quantal Theories of

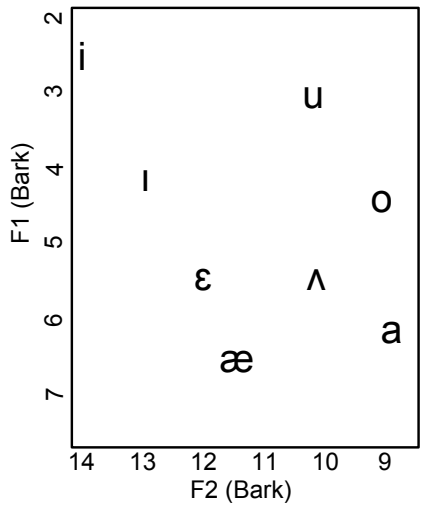


Figure 5.2: Mean formant frequencies of vowels for male speakers of California English in this experiment.

Vowel	**Acoustic Density**	**Lip Rounding**	**Palatal contact**
ʌ	High	No	Minimal
ɛ	High	No	Moderate
u	Low	Yes	Moderate

Table 5.2: Important characteristics of acoustic and somatosensory targets for three English vowels.

speech production (Stevens, 1972, 1989), vowels are located in stable acoustic regions where relatively large changes to articulation result in only small acoustic changes. If true, vowels are located in *articulatorily* sparse neighborhoods.

These acoustic and somatosensory target characteristics are likely to affect subject responses to altered auditory feedback. Sensitivity to acoustic feedback would lead to less compensation in vowels with more phonological neighbors than those with fewer neighbors. Furthermore, any compensation in dense phonological regions ought to be *indirect*, with subjects changing their formant production in a way that circumnavigates neighboring vowel regions. In addition, sensitivity to somatosensory feedback would lead subjects to compensate less for vowels with more salient somatosensory feedback. Predicted effects on the vowels tested here are set out in Table 5.3.

Because properties of acoustic neighbors make different predictions about the relative magnitude of compensation in the three test vowels than does salience of somatosensory feedback, comparing compensation among these three vowels will allow the evaluation of whether either sort of phonological knowledge is involved in

Vowel	Acoustic: Mag.	Acoustic: Direct.	Somato.: Mag.
ʌ	Small	Direct	Large
ɛ	Small	Indirect	Medium
u	Large	Direct	Medium

Table 5.3: Predicted influence of acoustic and somatosensory information on compensation across the vowel space.

compensation for altered auditory feedback.

Methods

Subjects (n=20, all males) sat in a sound booth. All subjects reported that they were native speakers of English and had normal hearing and language skills. They were paid a small sum for their participation. Participants were assigned to one of three groups: the /ɛ/ condition, the /ʌ/ condition, or the /u/ condition. They used the Feedback Alteration Device (FAD) described in Chapter 3, which involves a headset wired so that microphone input is fed through a computer and back into its earphones in real time. During three sessions on three separate days, they produced isolated monosyllabic words under the following conditions:

/ʌ/ and /u/ conditions

1. Control. Subjects produced {b,c,d}Vd words presented visually on the screen (e.g., *bed*, *bad*, *dad*, *cod*). Real time feedback was played back through the earphones but was not altered.

2. F2 UP condition. One of the stimulus words was generated at random and displayed on a computer screen a total of 200 times, once every 2 seconds. Subjects were instructed to produce the word when it appeared. Each subject's F2 feedback was altered slowly from 0 Hz up to the maximum F2 shift over the course of the experiment.

3. Complementary condition. One of the stimulus words was generated at random and displayed on a computer screen a total of 200 times, once every 2 seconds. Subjects were instructed to produce the word when it appeared. /u/ subjects heard their F2 feedback altered slowly down to a maximum shift of -300 Hz, and /ʌ/ subjects heard their F1 feedback altered slowly up to a maximum shift of 250 Hz.

The complementary condition was designed to answer a different experimental question and will not be analyzed here.

Phase	# Trials	Formant shift
1	15	No shift
2	65	F2 is shifted in even increments up (or down) to the maximum shift.
3	90	Maximum F2 shift
4	30	No shift

Table 5.4: Experimental design, /ʌ/ and /u/ conditions.

/ɛ/ condition

The design of this experiment was slightly different for the /ɛ/ experimental subjects. Subjects produced only the word 'head' during the control condition, and they produced 360 words on each day, not 200. The procedure for participants in the /ɛ/ condition was as follows:

1. Control. The word 'head' appeared as a prompt on a computer screen a total of 360 times, once every 2 seconds. Subjects were instructed to produce the word when it appeared. Real time feedback was played back through the earphones but was not altered.

2. HEAD F2 condition. The word 'head' appeared as a prompt on a computer screen a total of 360 times, once every 2 seconds. Subjects were instructed to produce the word when it appeared. Each subject's F2 feedback was altered stepwise from 0 Hz to 250 Hz over the course of the experiment. Feedback was altered in 5 "steps": a 45-trial ramp and then a hold at a 50 Hz shift, followed by another 45-trial ramp up to a hold at a 100 Hz shift, a third ramp up to a hold at 150Hz, a fourth up to 200 Hz, and a fifth up to 250 Hz. The sequence of formant shifts over the course of the experiment is shown in Figure 5.3.

3. HEAD F1 condition. On a third day, subjects participated in a complementary condition in which their F1 was altered stepwise from 0 to 250Hz, using the same experimental design as the HEAD F2 condition. This condition was addressed in Chapter 3 and will not be discussed here.

For the /ɛ/ condition, all stimulus words were 'head'. For the /ʌ/ condition, stimulus words were 'bud', 'dud', and 'hud', and in the /u/ condition, stimulus words were 'bood', 'food', and 'rude'[4].

It is important to note that these three experiments have methodological differences that makes it difficult to compare the outcomes of the experiments. The major difficulties are twofold: first, that the maximum amount of formant shift is different

[4]These /u/ words, 'bood', 'food', and 'rude', were chosen because, for most speakers, 'bood' and 'food' vowels fall in the middle of the /u/ vowel region, and 'rude' falls slightly further front.

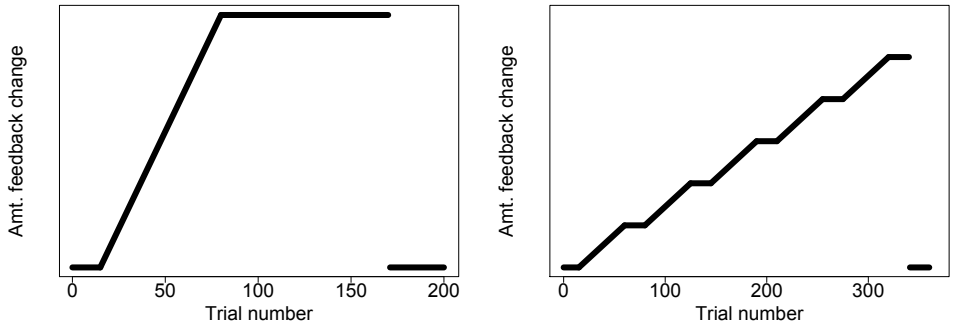


Figure 5.3: (Left) Change in feedback over the course of the 200-trial (/ʌ/ and /u/) experiments. Both experiments began with a 15-trial period of veridical feedback, followed by a 65-trial ramp up to the maximum feedback shift. For the last 30 trials, feedback was unshifted. (Right) There were 360 trials in the /ɛ/ experiment. These 360 trials were composed of 6 regions of equal formant alteration (plateaus) connected by ramps of slowly increasing feedback alteration.

for each experiment, and second, that the /ɛ/ experiment has only one stimulus word, while the other two experiments have three. If the maximum amount of compensation is dependent on the size of the formant shift, as suggested in Chapter 3, then the three experiments will have different compensation ceilings. If the number of stimulus words affects the amount of compensation, then behavior in the /ɛ/ condition will differ from the other two conditions for reasons unrelated to the salience of feedback in the 'head' vowel. These difficulties will be addressed in the discussion section.

Results

As explained in the previous chapters, most published analyses of compensation data analyze F1 and F2 compensation separately. However, Chapters 3 and 4 (along with a host of previous literature) demonstrate that speakers do not manipulate F1 and F2 independently, and for this reason it is worth analyzing F1 and F2 jointly. Both types of analyses are presented here and yield their own insights into compensation for altered auditory feedback. Separate formant analyses are presented first.

When F1 and F2 are measured separately, subjects compensated or marginally compensated for the feedback shift in all three vowels. Figure 5.4 shows subject responses to the shift in F2 feedback for /ɛ/. To normalize across subjects, each subject's average F1 and F2 during the first 15 (no-shift) trials was subtracted from all of his formant measurements. For example, a normalized measurement of 40 Hz would indicate that a subject produced a formant 40 Hz above his baseline average.

The blue dots in these graphs show the normalized formants that subjects heard at each trial, and the black dots show the normalized formants that they produced at each trial. Each dot represents the average across all subjects in that condition at that trial. In the bottom graph, which shows F2 over the course of the experiment, the blue and black dots are on top of each other for the first 15 (no-shift) trials. Starting at trial 16, the blue and black dots diverge until, at trial 320, they are 250 Hz away from each other. Notice that as formant feedback begins to shift, subjects begin to produce vowels with lower F2: they compensate for the change in feedback. In the top graph, which shows F1 over the course of the experiment, the blue dots and the black dots are in roughly the same place, indicating that subjects heard the same formants that they produced. Surprisingly, even though auditory feedback was not shifted, subjects increase the F1 they produce over the course of the experiment.

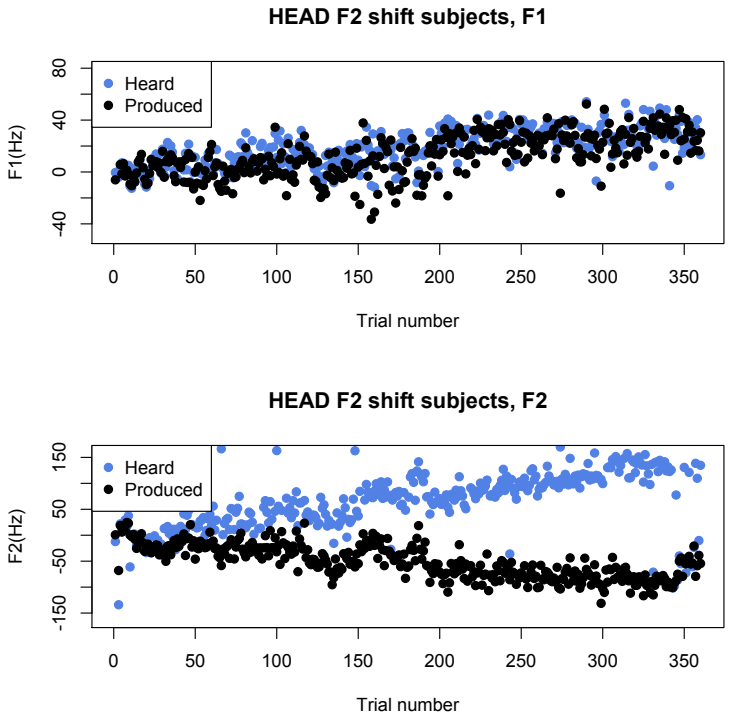


Figure 5.4: Response to /ɛ/ F2 feedback shift. Blue dots show the formants that subjects heard at each trial, and the black dots show the formants that they produced at each trial. Each dot represents an average across all /ɛ/ subjects.

Figure 5.4 shows that subjects in the /ɛ/ condition responded to the F2 feedback shift, on average, with a change in F2 production of -91.3±11.9Hz, or 36% of the expected F2 shift. They also changed their F1 production by 29.2±10.0Hz. Although there is clear compensation across subjects, some compensated more than others. To

account for differences in compensation among this small group of subjects, a mixed-effects analysis with no fixed effect (except an intercept) and subject as a random effect is used to measure compensation at trials with maximum shift. This analysis determines whether F1 or F2 compensation is significantly different from 0 when each subject is allowed to have his own mean compensation. As in the previous mixed-effects analyses, the intercept is estimated using the `lmer` function from the R package `lme4`, and confidence intervals for the intercept estimate is found using the function `pvals.fnc` from the R package `languageR`. Results are shown in Table 5.5.

Formant	2.5% Estimate	Mean Estimate	97.5% Estimate	$Pr(>\mid t \mid)$
F1 intercept	**15.2**	**30.3**	**45.9**	**0.004**
F2 intercept	**-63.2**	**-89.1**	**-113.5**	**0.0005**

Table 5.5: Summary of mixed-effect linear regression model estimating F1 and F2 compensation for /ɛ/ at maximum shift with subject as a random effect. Compensation is significantly different from 0 in both F1 and F2.

Figure 5.5 shows that subjects compensated for the /ʌ/ feedback shift as well. Again, to normalize across subjects, each subject's average F1 and F2 during the first 15 (no-shift) trials was subtracted from all of their formant measurements.

On average, during the last 40 trials with maximum shift, subjects compensated -64.9±24Hz in F2 (or 16% of the 400 Hz feedback shift), and 0±9.8Hz in F1. A mixed effects analysis modeling F1 or F2 compensation with subject as a random effect fails to show significant compensation in either F1 or F2. Confidence intervals for the two analyses are listed in Table 5.6. The marginal compensation in F2 suggests that there is substantial individual variation in compensation for altered auditory feedback. And indeed, one subject in this experiment failed to compensate, and one followed the formant shift.

Formant	2.5% Estimate	Mean Estimate	97.5% Estimate	$Pr(>\mid t \mid)$
F1 intercept	-17.7	-0.13	18.1	0.99
F2 intercept	-106	-65.2	-23.3	0.07

Table 5.6: Summary of mixed-effect linear regression model estimating F1 and F2 compensation for /ʌ/ at maximum shift with subject as a random effect. Compensation is not significantly different from 0 in F1 and marginally different from 0 in F2.

Figure 5.6 shows that subjects in the /u/ condition compensated more than subjects in the other two conditions. Again, measurements were normalized by subtracting each subject's average F1 and F2 during the first 15 (no-shift) trials from all of their formant measurements. For the last 40 maximum shift trials, /u/ subjects

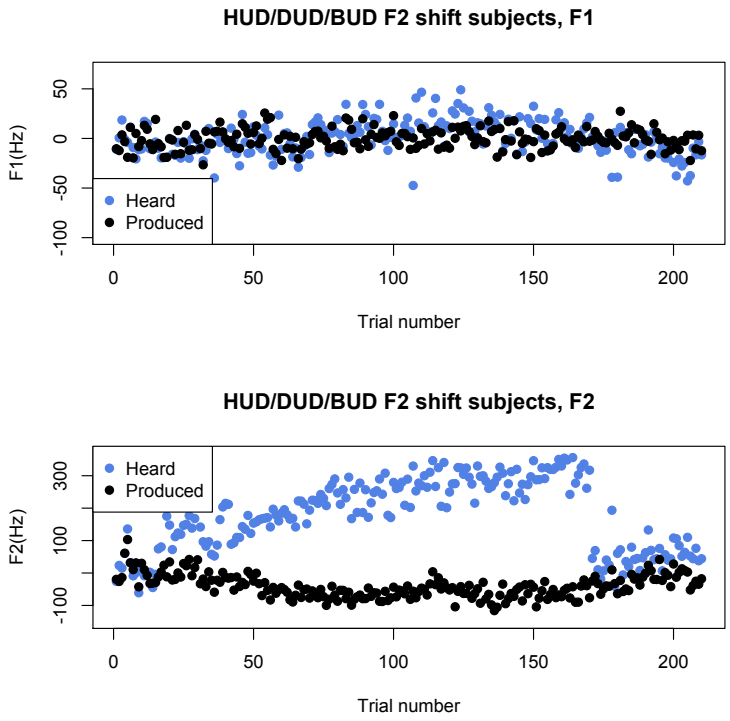


Figure 5.5: Response to /ʌ/ F2 feedback shift. The x-axis marks the trial number. Blue dots mark the normalized formant that the talkers head, on average, at each trial. Black dots show the formants they produced at each trial, on average. The top graph shows F1 over the course of the experiment, and the bottom graph shows how subjects compensate for a shift in /ʌ/ F2. See text for normalization procedure.

compensated in F2 by -112.4±61.8Hz, equivalent to 37% of the 300 Hz feedback shift, but only by 0.6±20.6Hz in F1.

A mixed-effects analysis estimating F1 or F2 compensation at maximum shift, with subject as a random effect, confirms that compensation is significant in F2. Confidence intervals for this analysis are shown in Table 5.7.

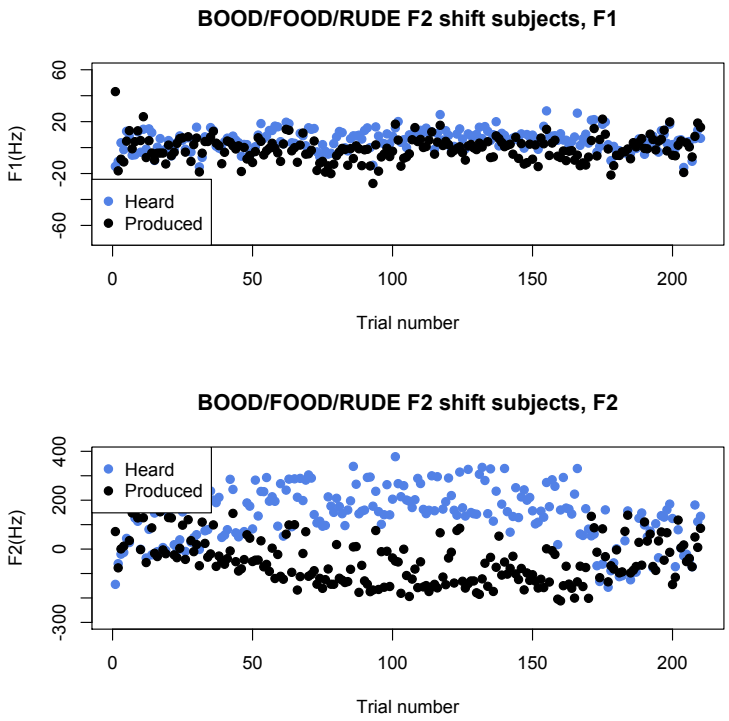


Figure 5.6: Response to /u/ F2 feedback shift.

Formant	2.5% Estimate	Mean Estimate	97.5% Estimate	$Pr(>\mid t \mid)$
F1 intercept	-10.9	-2.3	6.3	0.53
F2 intercept	**-192**	**-119**	**-46.0**	**0.003**

Table 5.7: Summary of mixed-effect linear regression model estimating F1 and F2 compensation for /u/ at maximum shift with subject as a random effect. Compensation is significantly different from 0 in F2, but not F1.

Though percent compensation was reported as the percent F2 production difference between no-shift trials and an F2 feedback shift, it is clear that many subjects compensate for altered feedback in F2 by changing their F1 *and* F2 production. If we measure percent compensation by measuring only the change in F2 production, we may underestimate the true production change. Thus, this second set of analyses explore F1 and F2 data from the 3 vowel conditions jointly. The first analysis measures the average magnitude of compensation using a wedge plot.

As explained in Chapter 4, wedge plots are essentially two-dimensional boxplots. They normalize formant measurements across subjects by converting them into polar coordinates: (1) *magnitude*, the distance between the center of the baseline region and the current (F1, F2) coordinate pair, measured as a Euclidean distance in Bark F1-F2 space, and (2) *angle*, a measurement of how much the subject compensated for feedback shifts in one formant with production changes in another formant. These plots show the range of formant production change exhibited across subjects during the trials with maximum formant shift. These plots have several advantages over boxplots and the time-course plots of F1 and F2 already presented. One advantage is that they can show normalized formant production across subjects with respect to adjacent vowels in vowel space. Another is that they report formant statistics characterizing all trials with maximum F2 shift, rather than reporting equivalent trials separately, as the time course plots do. Looking at the data this way takes into account that trials with maximum formant shift are meant to be equivalent with respect to the amount of shift and the expected response to that shift; trial 140 ought to be exchangeable with trial 160. The plots also show the variability in production across all subjects. The black dotted lines in this figure show the amount of formant shift in each condition, and the gray dotted lines show the compensation expected if the subject opposed the formant shift directly and completely.

The wedge plots in Figure 5.7 show that subjects clearly responded to the feedback shifts in all three conditions. Wedges display the 25th and 75th percentiles of magnitude and angle response to F2 feedback shift in /ɛ/, /ʌ/, and /u/. The black dotted line shows the amount of formant shift – the mean vowel that subjects would have heard had they not compensated at all. The grey dotted arrow shows "perfect compensation" – the mean vowel that subjects would have to produce at maximum shift to hear their mean baseline vowel formants. These plots show that the amount of production change did not scale either with the amount of shift or with the degree of auditory or somatosensory target interference. The smallest feedback shift, in /ɛ/, coincided with substantial interference from phoneme neighbors and moderate somatosensory feedback. The amount of production change was larger than for /ʌ/, even though /ʌ/ had the largest shift in F2 feedback and the smallest amount of interference from somatosensory feedback and phoneme neighbors. Also unexpected was the large response to the /u/ F2 shift, especially given its salient lip rounding feedback. These findings are scrutinized further in the interim discussion.

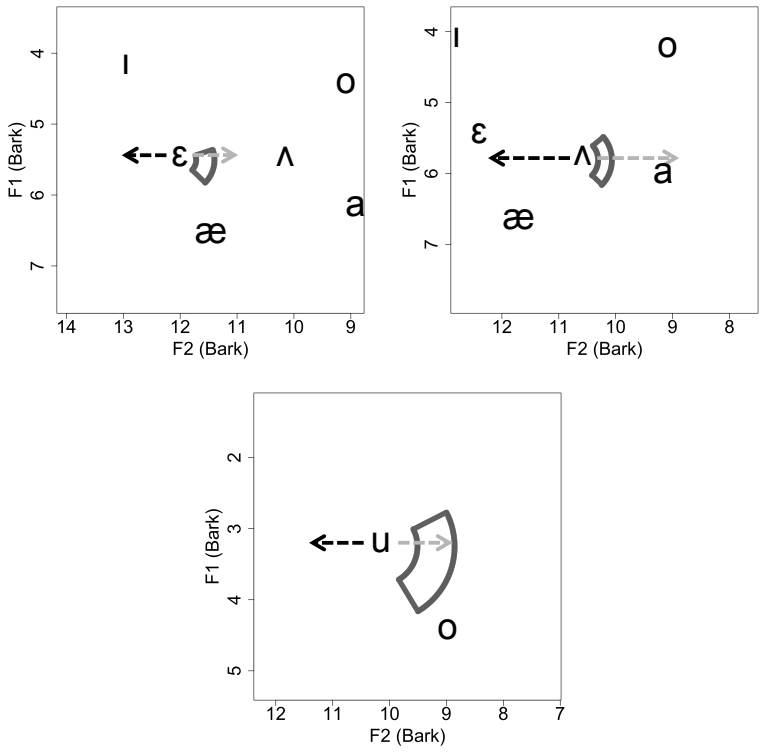


Figure 5.7: Wedges display the 25th and 75th percentiles of magnitude and angle response to F2 feedback shift in /ɛ/, /ʌ/, and /u/. (Top left) The middle 50% of formants produced across subjects in response to a 250Hz shift in F2 /ɛ/ feedback. (Top right) The middle 50% of formants produced across subjects during the last 40 trials with 400Hz shift in the F2 of /ʌ/. (Bottom) The middle 50% of formants produced across subjects during the last 40 trials with 300Hz shift in the F2 of /u/. Formants were normalized with the magnitude-angle method described above.

In addition, these plots make it clear that subjects often failed to oppose the feedback shift directly: although the feedback shift operated only on F2, subjects often changed their production of F1 as well. The range of angles was fairly well centered for shifts in /ʌ/, but leaned heavily toward a higher F1 in /ɛ/ and /u/.

The average magnitudes and angles of compensation across subjects is listed in Table 5.8. On average, subjects responded much more to shifts in /u/ than to shifts in either /ɛ/ or /ʌ/. However, there is substantial variation in both magnitudes and angles across subjects. While some opposed the feedback shift swiftly and directly, others wandered aimlessly around their baseline vowel spaces. The large standard deviations on magnitudes and angles make this point quantitatively.

So far, results of this experiment have demonstrated that subjects adjust their articulatory plans when there is a mismatch between observed and expected F2 audi-

Vowel	Magnitude	Heard mag.	% Comp.	Angle
/ɛ/ F2	0.53±0.21	0.80± 0.13	66%	27.0±50.2°
/ʌ/ F2	0.50±0.28	1.59± 0.41	31%	-3.9± 70.2°
/u/ F2	0.92± 0.66	1.61± 1.19	57%	15.8± 66°

Table 5.8: Average magnitudes and angles of compensation across subjects for the three altered vowels.

tory feedback in three different regions of vowel space. But by themselves, changes in production do not reveal whether the phonological inventory influences the comparison between observed and expected feedback. Now that it is certain that subjects are compensating, the next step is to examine the direction of compensation and evaluating whether subjects seeks to avoid *saying* or *hearing themselves say* adjacent phonemes. Recall that the experimental design creates a mismatch between the two: what subjects say is not what they hear themselves say.

To determine whether subjects are sensitive to hearing adjacent phonemes or saying adjacent phonemes, the augmented wedge plot in Figure 5.8 shows the mean formants that subjects heard during the maximum F2 feedback shift in each of the three conditions. If subjects were compensating completely, the green "x" would fall on top of the baseline /ɛ/, /ʌ/, or /u/. Because they are not compensating completely, we can observe whether they seem to be producing vowels that help them avoid *hearing* adjacent phonemes.

Figure 5.8 shows that, on average, subjects hear vowels that are relatively far from the center of their baseline vowel regions by the time their feedback has reached the maximum formant shift. Subjects do not appear to change their vowel production so as to avoid hearing adjacent vowels. One one hand, subjects never actually produce adjacent vowels. On the other hand, only the /ʌ/ condition put subjects in a situation where failing to compensate would cause subjects to hear a competing vowel. Subjects in this condition compensated enough, on average, to avoid hearing a vowel at the center of /ɛ/ space, but multiple subjects still ended up hearing a vowel confusable with /ɛ/. Again, the black dotted lines in this figure show the amount of formant shift in each condition, and the gray dotted lines show the compensation expected if the subject opposed the formant shift directly and completely.

Interim discussion

Compensation for altered auditory feedback seems to depend on the altered phoneme's phonological neighborhood, highlighting the importance of auditory targets in vowel representations. As predicted, compensation for /ʌ/ and /ɛ/, which are located in dense phonological regions, is small in magnitude, and compensation for /u/, which is located in a sparse phonological region, is large. Observed and ex-

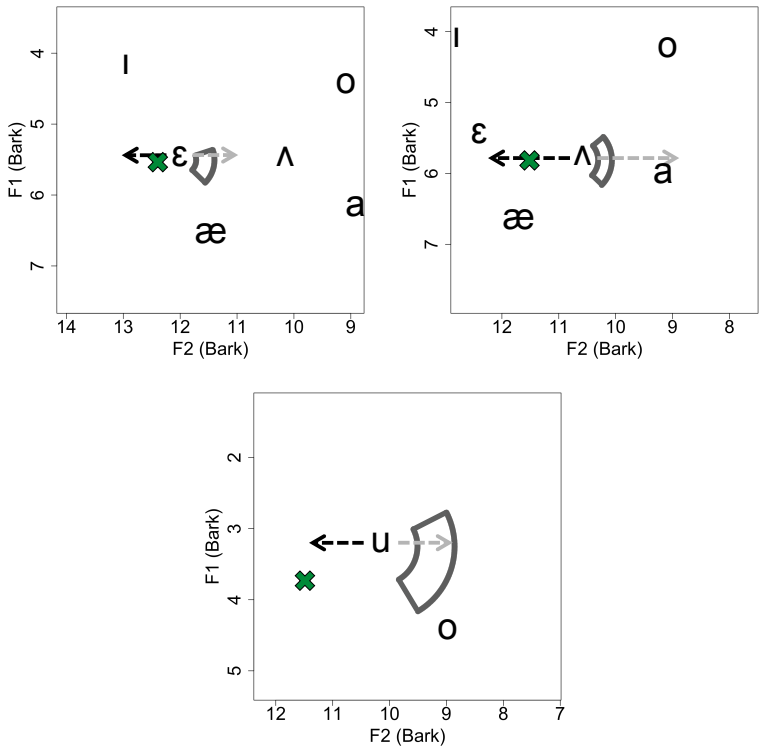


Figure 5.8: Green "x" shows the mean formants heard during the maximum F2 feedback shift in /ɛ/ (top left); /ʌ/ (top right); /u/ (bottom). Formants were normalized with the magnitude-angle method described above. Wedges display the 25th and 75th percentiles of compensation magnitudes and angles during these trials.

pected results for all three vowels are summarized in Table 5.9. Phonological density predicts the absolute magnitude of compensation very well. However, phonological density does not match as well with percentage compensation, which normalizes compensation magnitude for the amount of shift. This is likely due to the compensation limit observed in Chapter 3. Because compensation is more complete for smaller feedback shifts, compensation ought to be most complete for the smallest shift and least complete for the largest shift. This experiment shows that the shift size effect influences compensation independent of vowel density. Percent compensation is greatest for the 250Hz shift in /ɛ/ and least for the 400Hz shift in /ʌ/, with percent compensation for /u/ falling in between.

Somatosensory targets, by contrast, cannot explain the cross-vowel responses found in this experiment. If somatosensory salience were driving compensation for altered F2 feedback, then /u/ compensation would be smaller in magnitude than /ʌ/ or /ɛ/ compensation. Instead, percent compensation turned out to be *greater* for /u/ than for the other two vowels, and subjects compensate least for /ʌ/, which had the

largest feedback shift and the *least* salient somatosensory feedback.

The experiments in this chapter support a model of speech motor control that accesses phoneme representations when creating new articulatory plans.

Vowel	**Acoustic: Mag / Angle**	**Somato.: Mag**	**Actual results**
/ʌ/	Little / Direct	Lots	0.50Bark / 31% of F2 shift / Direct
/ɛ/	Little / Indirect	Some	0.53Bark / 66% of F2 shift / Indirect
/u/	Lots / Direct	Some	0.92Bark / 57% of F2 shift / Indirect

Table 5.9: Predicted and actual effects of acoustic and somatosensory information on compensation across the vowel space.

The phonological neighborhood effect is described as avoiding producing adjacent vowels, yet in no case do talkers produce formants anywhere near other vowels. In order for this effect to plausibly be a phonological density effect, talkers would have to know how any potential set of formants would be categorized, and aim to avoid not only prototypical exemplars of competing vowels, but also intermediate vowel sounds confusable with adjacent vowels. In order to avoid adjacent phoneme *regions*, a talker's speech motor control system must have access not only to information about prototypes or centers of other vowels, but also to a acoustic map of "hot spots" and "cold spots" – with labels – for the talker's entire vowel space.

Although this experiment tested for effects of auditory vowel density, it did not test for effects of articulatory vowel density. Evidence that the phonological inventory affects adjustment for altered auditory feedback could come in two forms: subjects either aim to produce vowels that *sound* like the intended vowel when compensating for altered auditory feedback, or aim to produce vowels that *feel* like the intended vowel when they compensate. A well-designed compensatory mechanism might save effort by amending motor plans with corrective utterances that weren't entirely new. Such a mechanism would predispose talkers to say things that they have said before.

A tendency to produce vowels that have been produced before can be viewed as a frequency effect. Frequency effects appear often in the cognitive science and psycholinguistics literature. There is ample evidence for a perceptual magnet effect (Kuhl, 1991), wherein talkers are predisposed to hearing sounds that they have heard frequently before. It is also known that well-practiced movements are faster to produce. Could there also be a 'production magnet effect', where if we intend to produce an infrequent vowel, we naturally gravitate toward producing a similar-sounding frequent one?

Subjects' articulation was not measured during the experiment, and it is difficult

to infer the articulations they used to compensate, but the vowels they heard were recorded. Based on these recorded vowels, shown in Figure 5.8, subjects appear to avoid producing rather than hearing adjacent vowels. Compensation magnitude and angle was unrelated to the positions of adjacent vowels. Taken together, the perception and production evidence suggest that the phonological inventory influences the adjusted articulatory plan derived from the mismatch between observed and expected auditory feedback. If the comparison between observed and expected feedback was itself influenced by the phonological inventory, subjects would aim to avoid hearing themselves say adjacent phonemes, and there is no evidence that they do this. The schematic of speech motor control presented at the beginning of this chapter should be updated as in Figure 5.9.

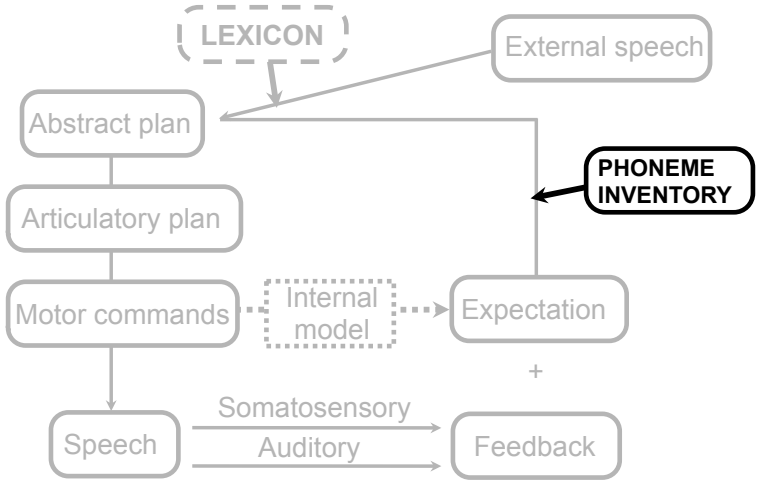


Figure 5.9: Amended model of speech motor control.

Still, it is impossible to dismiss the idea that small differences in the designs of these three experiments are driving the differences in response between them. For example, Chapter 3 showed that percent compensation decreases with increasing formant feedback shift. For that reason, we should expect that percent compensation for the 400 Hz shift in /ʌ/ should be smaller in magnitude than for a 250 Hz shift in /ɛ/. But even the raw magnitude of compensation for /ʌ/ is slightly smaller than for /ɛ/, and much smaller than for /u/. The magnitude of compensation remains surprisingly small for /ʌ/.

The /u/ results are broadly consistent with a production magnet effect. When the /u/ in 'bood', 'food', or 'rude' is shifted up or down by 300 Hz, the resulting vowel is still a convincing rendition of /u/, in part because /u/ occupies such a large region of formant space in California English. The vowel talkers hear when 'bood' is shifted is familiar, and the vowel associated with complete compensation is also a familiar /u/.

The observed results are also consistent with an articulatory frequency effect. That is, if frequency effects warp the space of possible solutions to mismatched feedback, we might expect talkers to compensate by producing frequently-produced vowels, even if those vowels don't exactly oppose the formant shift. The place where this might be most visible is along the edges of vowel space, where production is constrained by physiological limits. It is not possible to produce a midvowel front of [ɛ] or a low vowel front of [æ]. Accordingly, talkers might be especially inclined to compensate for an F1 shift in [ɛ] or [æ] by changing their F1 and F2.

This is indeed what we find. Frequent vowels fall along a diagonal in the vicinity of /ɛ/, and compensation for this condition involves a greater F1 component than the other conditions. Of course, this may either happen because the vowels along the vowel periphery are more frequent than the vowels required to oppose the feedback shift, or because subjects hear the change as a combination of formants.

If a vowel's token frequency is critical to compensation for altered auditory feedback, talkers ought to compensate for small, within-category changes in other vowels as well. A follow-up experiment tests the articulatory familiarity hypothesis by altering /ɛ/ by a two much smaller amounts, 30Hz and 90 Hz[5].

Two small-shift experiments

Two small formant feedback shift experiments demonstrate the character of compensation for subcategorical changes in feedback. These experiments measured the response to 30 and 90Hz /ɛ/ feedback shifts in F1.

There were 7 male subjects in this experiment, all native speakers of California English. One subject was excluded from most analyses due to technical difficulties when collecting his baseline vowel space.

Procedure: The word 'head' was displayed on a computer screen once every 2 seconds. Subjects were instructed to say 'head' loudly and clearly each time that it appeared on the screen. They were recorded while speaking. The sequence of formant shifts over the 360 trials is detailed in Table 5.10.

The vowel formant measurement procedure was the same as in the other experiments. It used a perl script running Entropic's ESPS/Xwaves utilities to find formants in the middle 50ms of each vowel.

Results: Figure 5.10 shows that subjects compensate for this small shift in /ɛ/ F1 feedback. As in the other experiments presented in this chapter, this figure illustrates compensation with formants that have been normalized across subjects. Just as in the other experiments, normalized formants were computed by subtracting each

[5]Even with altered auditory feedback, frequently produced vowels have frequently-accessed motor plans and frequently-felt somatosensations. To distinguish these results from a "somatosensory familiarity" hypothesis, one could repeat this experiment in subjects whose mouths have been numbed with a topical gel.

Phase	# Trials	Formant shift
1	75	No shift
2	30	F1 is shifted an additional 1 Hz per trial up to 30 Hz.
3	90	30 Hz F1 shift
4	40	F1 is shifted an additional 1.5 Hz per trial up to 90 Hz.
5	90	90 Hz F1 shift
6	35	No shift

Table 5.10: Design of small F1 shift experiment.

subject's average F1 and F2 during the first 75 (no-shift) trials from all of their formant measurements. The blue dots in these figures show the normalized formants that subjects heard at each trial, and the black dots show the normalized formants that they produced at each trial. Each dot represents the average across all subjects in that condition. In the top graph, which shows F1 over the course of the experiment, the blue and black dots are on top of each other for the first 75 (no-shift) trials. Starting at trial 76, the blue and black dots diverge until, at trial 245, they are 90 Hz away from each other. Notice that as formant feedback begins to shift, subjects begin to produce vowels with lower F1: they compensate for the change in feedback. Surprisingly, even though F2 auditory feedback was not shifted, subjects increase the F2 they produce over the course of the experiment.

This data gives us a set of formant values produced with no shift and a set of formant values produced with a 30 Hz shift and a 90 Hz shift. Compensation has occurred if the latter two sets of formants are different from the no-shift set. This amounts to determining whether the *distribution* of formants produced with no shift is different from the distribution of the set of formants produced under either feedback shift. The three distributions are shown in Figure 5.11.

Clearly, there is some overlap between the three distributions, but there are also substantial differences between them. Baseline formants appear to be centered around 0, whereas the formants produced under a 30Hz and 90Hz shift look to be skewed left. If compensation has occurred, there should have been a significant and substantial change in the location or shape of the formant distribution between baseline and either shift. Due to the skewing, such a change is not appropriately diagnosed with a t-test[6]; the data are not normally distributed. Instead, a nonparametric test that detects changes in either the location and the shape of two underlying distributions, the Kolmogorov-Smirnov test, is used. This test yields a p-value indicating whether the two data sets are likely to have come from the same distribution, and a statistic

[6]A t-test does find that the formants produced at the 30Hz and the 90Hz shifts are both significantly different from the baseline formants, $p < 0.01$)

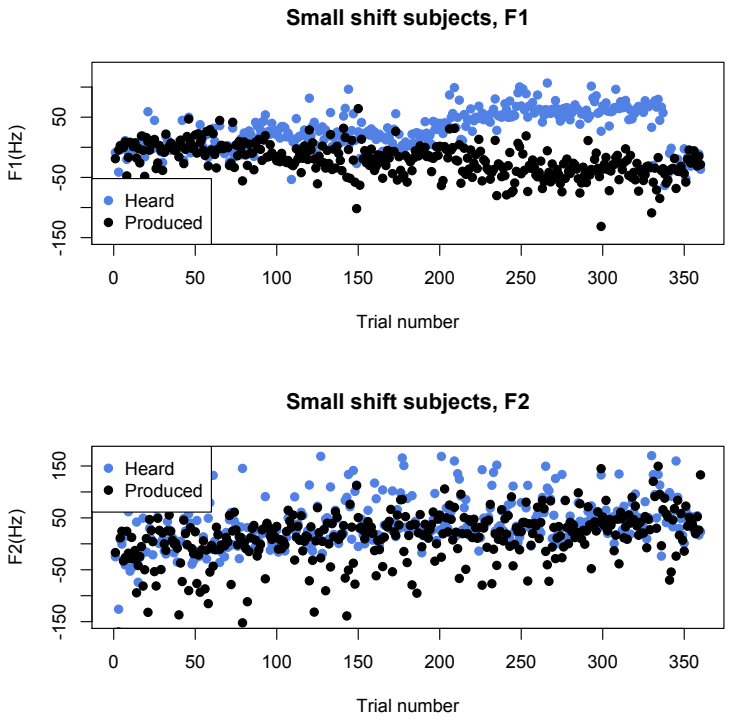


Figure 5.10: Response to 30 and 90Hz /ɛ/ F1 feedback shift.

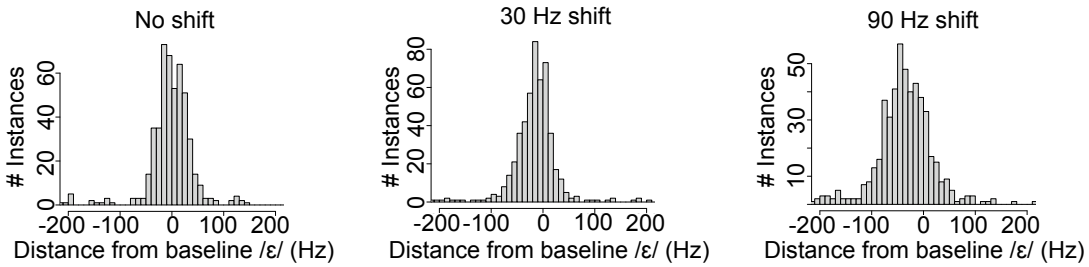


Figure 5.11: Response to 30 and 90Hz /ɛ/ F1 feedback shift. Leftmost graph: Normalized formants produced under no formant feedback shift. Center graph: Normalized formants produced with a 30Hz feedback shift in F1. Rightmost graph: Normalized formants produced with a 90Hz feedback shift in F1.

measuring the distance between the two distributions. Technically, the Kolmogorov-Smirnov test cannot be computed if there are identical values in this data (called "ties"), and because formant measurements from ESPS are computed as whole numbers, there are several ties in this data set. Ties were eliminated by adding a very

small amount of random noise (on the order of 0.001Hz) to the data.

This test, the results of which are shown in Table 5.11, confirms that the baseline formants are significantly different from both the formants produced with a 30 Hz F1 shift and formants produced with a 90 Hz F1 shift (p<2e-10).

Control	**Treatment**	**D**	**p-value**
No-shift	30 Hz shift	0.216	1.111e-10
No-shift	90 Hz shift	0.423	< 2.2e-16

Table 5.11: Kolmogorov-Smirnov tests for 30 Hz and 90 Hz shifts in the 'head' vowel.

The Kolmogov-Smirnov test confirms the intuition gathered from the raw data; talkers do compensate for very small shifts in formant feedback. They clearly have expectations for what they ought to sound like that are much more specific than the target region for either the word or the syllable – otherwise they would not be able to compensate for a formant shift this small. Additionally, compensation may be more complete for the 30Hz shift than for the 90Hz shift (based on F1 data). Though the variance of the data is large, the median of the 30Hz shift productions is -14Hz, or 47% of the 30Hz shift, and the median of the 90Hz shift productions is -36Hz, or 40% of the 90Hz shift.

Even though both of these shifts were quite small, the range of productions of 'head' was also relatively small. Many of the 90 Hz shifts bordered or fell slightly outside of that talker's baseline 'head' region. However, 30Hz shifts fell within this region. Appendix A shows convex hulls of shifted formants against baseline formants for each individual subject.

The 30Hz results clearly show compensation for a within-category shift. This result is all the more surprising because subjects show compensation below the 76 Hz compensation threshold proposed in Purcell & Munhall, 2006a, and because these formant differences are just barely discriminable, according to Kewley-Port, 2001.

A wedge plot of these results, shown in Figure 5.12, verifies that compensation is relatively complete, with high compensation variance.

Figure 5.12 indicates that, although subjects respond substantially in F1, they also produce changes almost as large in F2. Like the other /ɛ/ experiment, and unlike the /ʌ/ experiment, subjects seem to be responding to a feedback shift in F1 by recruiting both F1 and F2. Because this does not happen in all parts of the vowel space, it is unlikely that subjects are simply hearing all of their vowels as a combination of F1 and F2. Instead, a familiarity argument is more likely: subjects prefer to produce vowels that they have produced before. Perhaps altered phonemes that fall within the same category are faster and easier to interpret because the feedback signal matches the expected category, and it is more likely that the compensatory plan system knows what to do with such a sound. This makes subjects compensate more in vowels with large target regions, and compensate indirectly in areas of vowel

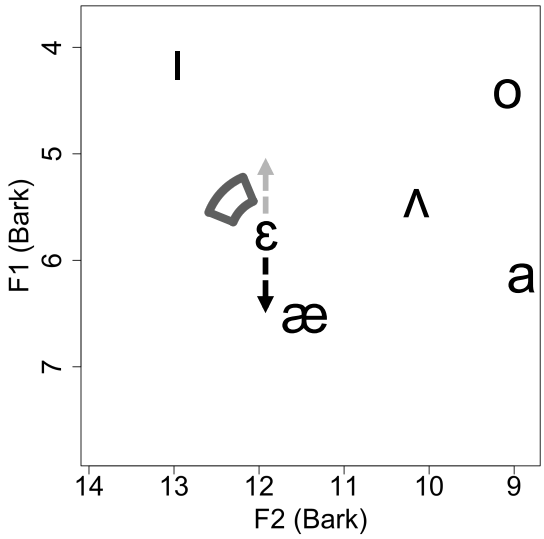


Figure 5.12: Wedges display the 25th and 75th percentiles of magnitude and angle response to the maximum (90Hz) F1 feedback shift in /ɛ/ across subjects. Formants were normalized with the magnitude-angle method described earlier.

space where the most frequent vowels lie along a diagonal.

This result is incompatible with the current implementation of the DIVA model of speech production; in that model, vowels are considered to be "on target" until they leave the talker's vowel target region.

Discussion

There are (at least) three characteristics that vary for vowels in different parts of vowel space. First, all phoneme pairs have different somatosensory feedback. Second, phonemes are in vowel regions of different density, with more or fewer vowels nearby. Third, different phonemes are situated closer to or further from the edge of the vowel triangle, making it more or less natural to compensate directly for a change in auditory feedback. It is possible that any of these three characteristics of the phonological inventory constrain the compensation response. The first and third characteristics are language universal, but the second, vowel region density, is a property of a particular language's phonological inventory.

Beginning from the assumption that vowel targets have auditory and somatosensory components, based on the results reviewed in Chapter 3, this experiment manipulated auditory feedback in a variety of somatosensory environments and regions of vowel space to observe how salient information from acoustic or somatosensory prop-

erties of vowel categories impinges on compensation. There were two main results. First, compensation was more direct for feedback manipulations in /ʌ/ than in /ɛ/ or /u/. Subjects were more likely to respond to an auditory feedback change in /ʌ/ with a production change in F2 rather than a combination of F1 and F2, but not so in /ɛ/ or /u/. The /u/ results may have been influenced by technical issues with formant synthesis, which for many subjects shifted F1 by 50-100Hz when F2 was shifted.

The magnitude of compensation was greatest for /u/ and least for /ʌ/, even though /u/ has more salient somatosensory feedback than does /ʌ/. The response to the /ʌ/ and /ɛ/ shifts were similar. This may be an effect of their similar phonological region densities. The indirect compensation for /ɛ/ was predicted by the articulatory frequency hypothesis. These results suggest that the salience of somatosensory feedback has little discernible effect on compensation for altered auditory feedback, at least for vowels that are not located at the corners of vowel space. If somatosensory information is weighted differently for different vowels, that weighting is not based on the salience of information in those vowels.

On the other hand, phonological density may influence compensation. Density was largest for /ʌ/, with five adjacent vowels within a 2 Bark radius, followed by /ɛ/, with three adjacent vowels within 2 Barks. California English /u/ had only one adjacent vowel within 2 Barks. Compensation followed exactly this pattern, with the smallest compensation for /ʌ/, followed closely by /ɛ/. Compensation in the /u/ condition was, on average, nearly twice as large. However, these three experiments had maximum formant shifts of different sizes. Recall that subjects will not continue to change their vowel production indefinitely as the formant shift they hear increases; eventually they hit an asymptote, beyond which they are unwilling to compensate further (MacDonald et al., 2010). This asymptote seems to occur around a formant shift of 300Hz for /ɛ/. If we assume that talkers are close to their production change limit for the three experiments, which used maximum F2 shifts of 250Hz, 300Hz, and 400Hz, then the absolute magnitudes are directly comparable. But because we don't know for certain whether these experiments caused talkers to reach their compensation limits, this observation is suggestive rather than conclusive.

A second analysis compared the vowels that subjects heard during maximum formant shift trials, in addition to the vowels that the subjects produced during these trials. Although subjects avoided producing adjacent vowels, there was no evidence that they avoided hearing adjacent vowels. Apparently the speech motor control system is phoneme-sensitive when constructing new articulatory plans, but is less phoneme-sensitive when making the initial comparison between observed and expected auditory feedback. The amended speech motor control model appears in Figure 5.9.

The interim discussion sought to explain why /u/ compensation is so large. One possibility was the phonetic properties of the /u/ phoneme category in California English. Not only does California /u/ have only one adjacent phoneme within 2 Barks, but it also takes up a lot of real estate in vowel space: it is the label for both

front /u/ productions such as 'toot' or 'dude', and back /u/ productions such as 'who'd'. There are two reasons that compensation might be greater in such a large phoneme region. One, subjects might be more inclined to compensate when they hear their voices shifted to a sound within a vowel category than to a sound outside of a vowel category. Two, subjects might find it easier to compensate for a sound that they have produced and heard before than a foreign, out-of-category sound that they have never had to deal with.

To decide whether formant shifts in familiar parts of vowel space really confer a compensation advantage, a follow-up experiment shifted /ɛ/ in a new set of subjects by two small amounts, 30Hz and 90Hz, in F1. This shift is small enough that it falls within, or very close to, the baseline /ɛ/ region for all of the subjects for whom we have a background vowel space. Mean percentage compensation for this shift is greater than for any of the larger F2 shifts tested in this chapter, or for the larger F1 shift tested in Chapter 3.

Indirect compensation in both /ɛ/ conditions and strong compensation for altered F2 in /u/ is consistent with an effect of acoustic and articulatory familiarity. Perhaps talkers prefer to hear vowels that they have heard before, and to produce vowels that they have produced before. Perhaps altered phonemes that fall within the same category are faster and easier to interpret because the feedback signal matches the expected category, and the articulatory planning system is more likely to know how to handle it. Articulatory familiarity seems to predispose a talker to compensate more in vowels with large target regions, and compensate indirectly in areas of vowel space where the most frequent vowels lie along a diagonal in F1-F2 space.

It should be noted that all of the analyses in this chapter relied on the exchangeability of the last 40 trials with maximum shift. That is, in the 200 trial experiments, there should have been no systematic difference between any consecutive or nonconsecutive pair of trials between trial 130 and trial 170. By inspection, these trials appear to be roughly the same, but it is possible that there is some subtle time-dependence, even during periods of consistent formant shift. The analyses in the next chapter investigate this possibility.

In spite of clear differences between experimental conditions, a few caveats are in order. First, this hypothesis is at odds with findings of Niziolek (2010), who finds that within-category compensation is *less complete* than compensation for sounds that fall outside of one's vowel category. Reconciling differences in experimental methodology may help to account for the contrasting results in these two lines of research. Second, it is difficult to control for all differences between phoneme pairs, and therefore difficult to rule out all alternative explanations for observed compensation. Third, there is substantial individual variation in compensation behavior. Some subjects respond with a large shift in production, and some barely compensate at all. Some recruit unaltered formants heavily when they compensate, occasionally even compensating in the wrong direction, while others have a more straightforward response to altered feedback. Because every subject has a different vowel space, with different relative

spacing between vowels and different sizes and locations for vowel target regions, this is an issue that is best tackled on an individual level. Chapter 6 seeks to understand some sources of this intersubject variation.

Chapter 6

Experiment 3: Modeling individual variation in compensation

The preceding chapters have demonstrated that, while there are some clear and interesting patterns in compensation for altered auditory feedback, there is also substantial individual variation in performance. While Chapter 4 provided no evidence that the lexicon affects compensation, Chapter 5 showed that the position of a vowel in vowel space does have an effect: compensation magnitude was greater for the vowel /u/ than for /ɛ/ or /ʌ/. Based on these results, I speculated that the large compensation for /u/ is driven by the large size of /u/'s vowel space in California English: there is a huge range of high vowel variants of /u/ in different consonant contexts. Because of this, when /u/ is shifted by 300 Hz in F2, the shifted vowel is still heard as an /u/, albeit an /u/ that tends to be used in a different consonant environment. Likewise, the /u/ vowel that a subject would have to produce when compensating for a 300 Hz F2 shift would be a perfectly good production of /u/, though again an /u/ that would usually be used in a different consonant context.

If talkers do compensate more when the vowels they hear or produce during the experiment are familiar, then it follows that individual talkers might be more or less likely to compensate for any particular shift, depending on the vowel's range of variation. For example, a subject with a very small /u/ region might compensate very little if the experiment would make him hear or produce vowels that he never produces. This chapter investigates the hypothesis that a talker's language *background*, including native language and individual habits of production, influence compensation for altered auditory feedback. This exploration may help to explain why some subjects do not compensate at all for altered auditory feedback, while others compensate in directions that we would not expect.

There are two primary reasons that compensation might vary from individual to individual. The first is variation in perceptual boundaries. For some speakers, an /ɛ/ pronounced with an unusually high F1 will be confusable with an /æ/. For other speakers, an /ɛ/ altered in this way is not confusable with any other vowel

in their inventories, perhaps because their /æ/ is more centralized than their /ɛ/ (in addition to F1 differences separating the two vowels), or because their dialect does not contain /æ/ at all. Obversely, compensation might vary due to variation in production targets. For some speakers, producing an /ɛ/ with a higher F1 will feel to them like /æ/, while for other speakers, producing the same formant change might feel like a vowel in a different part of the /ɛ/ region, or in an unused portion of their vowel space. For these speakers, producing an /ɛ/ with a higher F1 may be difficult (because it is a novel sound) or impossible (due to quirks of their vocal tract and tongue shape). Both of these factors are strong candidates for influencing individual responses to altered auditory feedback.

Thus far, this thesis has been able to document that speakers vary widely in their compensation for altered auditory feedback, but it has not been able to explain why. The purpose of this chapter is to start asking why by systematically investigating the link between an individual's production and his experimental performance. It is important to note that there are no standardized methods for understanding the relationship between a talker's vowel space in a non-experimental setting and his vowel space in an experimental setting. The exploratory analyses presented in this chapter are meant to illustrate some interesting patterns, but are not meant to be the final word on this topic. Further exploration of this sort of data has the potential to explain some of the individual variation common in psycholinguistic and experimental phonetics experiments.

In the exploratory analyses presented in this chapter, speakers are assumed to aim for the following three goals:

1. Produce a vowel that sounds like the intended vowel.

2. Produce a vowel that feels like the intended vowel.

3. Produce a vowel that feels familiar.

Because vowel spaces for perception and production are dynamic and multidimensional, they can only be estimated at a particular point in time. As a proxy for familiarity or ease of production, *vowel density*, the number of vowels typically produced in a subregion of vowel space, is used. As a proxy for perception and production, the range of formants produced in spontaneous vowels is used. The rationale is that, in one's own speech, production and perception boundaries ought to mirror each other. That is, when a talker intends to say "eh" and produces a particular set of formants, that talker ought to hear those formants as /ɛ/. The set of formants that a talker produces when s/he intends to say /ɛ/ should be a subset of all of the formants s/he produces as /ɛ/. If anything, this method of representing the formants perceived as /ɛ/ is conservative, but potential issues arising from using this space for both perception and production will be addressed later in this analysis.

Background vowel spaces

To determine whether individual differences in production habits are visible in casual speech, individual vowel densities from an annotated corpus of spoken English were compared. This preliminary analysis uses the Buckeye corpus, a repository of 40 casual interviews with male and female speakers from the Columbus, Ohio area. Interviews lasted from 30-60 minutes (Pitt et al., 2007).

Vowel densities constructed for a few of these speakers are clearly different, even across thousands of vowel tokens. To compare vowel densities, a small subset of two female and two male interviews were selected from the corpus. A Praat script selected vowel tokens from the annotated interviews and measured their formants. Figure 6.1 shows a density plot of all vowels in the interviews, regardless of word class. The plot was created with R's `hist2d` function, which groups the formant measurements into equally-sized bins and colors each bin based on the number of vowels it contains. Black bins represent locations in vowel space where the subject did not produce any vowels. Red bins contain a few vowels, yellow bins contain a moderate number of vowels, and white bins contain many vowels.

Figure 6.1 demonstrates that 30-60 minute interviews yield a useful approximation of an individual's vowel space. There are some similarities among the four speakers. They all show extensive vowel reduction and production of schwa-containing words, and their vowel spaces are convex, without internal blank regions; there do not seem to be impossible formants within the vowel triangle. But aside from these global similarities, these speakers have substantially different vowel spaces, even aggregated over thousands of vowels. Subjects 16 and 19, shown in the top left and bottom left graphs, respectively, produce the greatest number of vowels in the area of /ə/ and /ɘ/. But unlike Subject 19, Subject 16 has a second hot spot near /ɛ/. The areas of highest density for the other two subjects are not at /ə/: Subject 18 produces the most vowels near /ɑ/, and Subject 33 produces the greatest number of vowels near /ɪ/. In addition to differences in areas of high density, the four subjects differ somewhat in the overall shapes of their vowel spaces. All four vowel spaces are roughly triangular, but Subject 18 produces more high back vowels than the other 3 subjects. Even the relative locations of vowels vary between speakers. The centers of /ɪ/, /ɛ/, and /æ/ are marked on the vowel chart; these vowels are approximately collinear in Subject 18, but not in the other three subjects.

Figure 6.2 demonstrates that there is even ambiguity in vowel production within a single speaker. This figure compares the formants of all of the vowels labeled as /ɛ/ and all of the vowels labeled as /æ/. There is substantial overlap between the two regions. For example, F1=650Hz and F2=1700Hz is frequently produced to represent either vowel.

Because a subject's casual speech background is so potentially useful in predicting compensation behavior, this chapter analyzes a mock interview recorded from a subject who also participated in compensation experiments. This subject's compen-

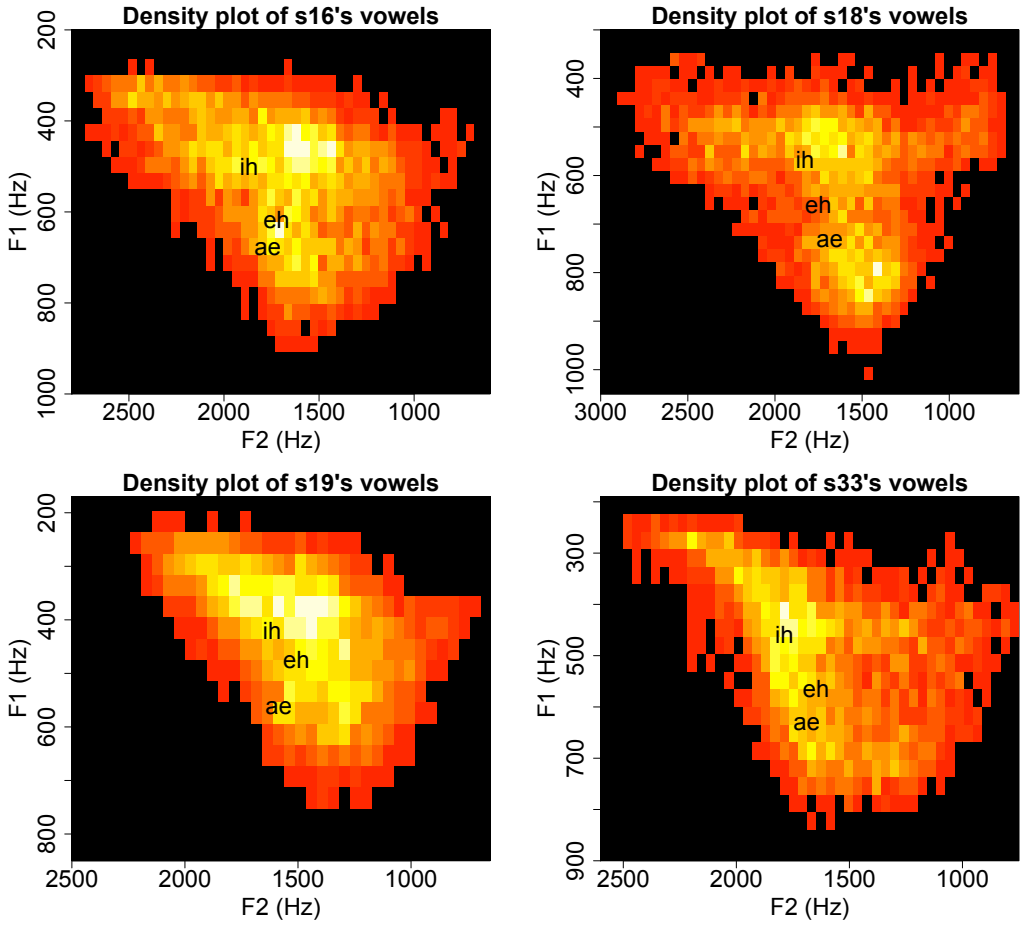


Figure 6.1: Vowel densities for 2 females and 2 males, taken from the Buckeye corpus. s16 and s18: old female; s19: old male; s33: young male.

sation data can be interpreted in light of his background vowel space. The hypothesis underlying the remainder of this chapter is that a talker's production experience, as indicated by a plot of vowel formants, may partly determine compensation to altered auditory feedback.

Data

Natural vowel space data. One male native speaker of California English participated in a half-hour mock interview covering neutral topics. His speech was recorded using an AKG Micro-Mic C520 microphone and a Marantz PMD660 Professional solid state recorder. The 4654 monophthongal vowels segmented from this interview were manually labeled using Praat and measured automatically with a Praat script.

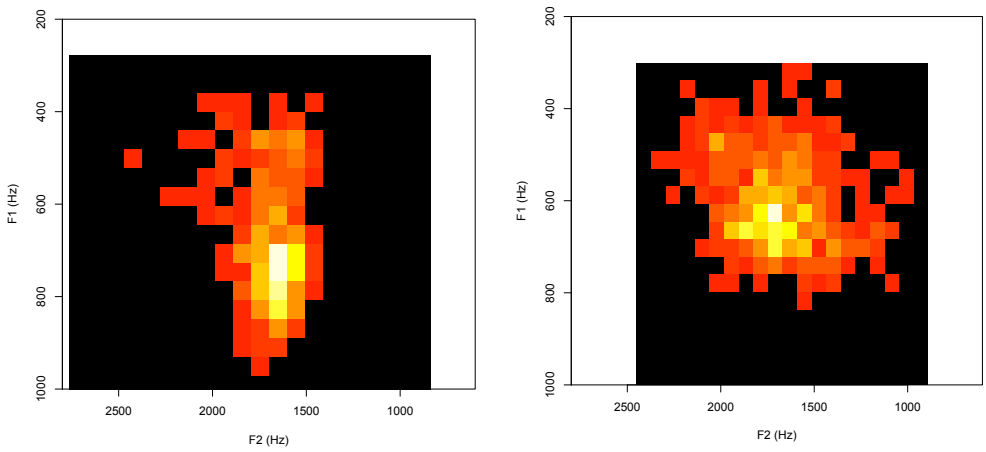


Figure 6.2: An older female's /æ/ and /ɛ/ vowel densities, respectively. Notice that there is a great deal of overlap.

The script used LPC, programmed to find 5 formants between 0 and 5000 Hz, to measure the average F1, F2, and F3 during the middle 50% of each vowel. Diphthongs and disfluencies were not measured, but all function and content words were included. Formant measurements were spot checked for outliers. As in the Buckeye corpus analysis, a density map was made from these measurements by dividing vowel space into $30 \cdot 30 = 900$ equally-sized bins, and counting the number of vowels from the mock interview that fell in each bin, regardless of label. Bins with higher counts were visited more often and bins with zero counts were never visited.

Experimental data. This subject also participated in 5 of the altered feedback experiments covered in the previous chapters:

1. /ɛ/ -250Hz F1 shift
2. /ɛ/ to /ɪ/ shift
3. /ɛ/ to /æ/ shift, linguistic version
4. /ʌ/ -250Hz F1 shift
5. /u/ +300Hz F2 shift

Prior to completing these 5 experiments, the subject also participated in a no-shift experiment with 360 trials containing a variety of monosyllabic CVC words. The control trial provides a second type of background vowel space, constructed from citation form vowels.

Figure 6.3 demonstrates that there are major differences between the casual speech density map and the citation form density map for the same individual. The

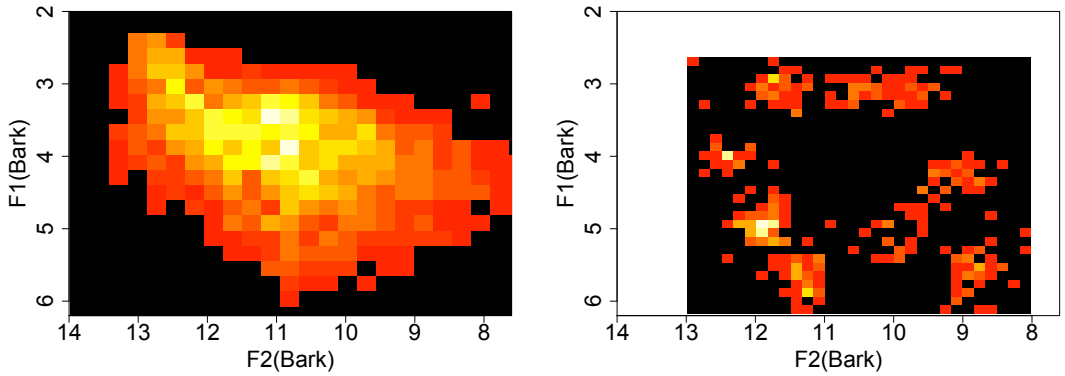


Figure 6.3: Vowel density map of casual speech from a male native English speaker, created from (Left) 4654 vowels extracted from a 30-minute mock interview, and (Right) 360 vowels in citation form CVC words. Vowels are not labeled in either map. Density is depicted in heat colors. The lighter the square, the more vowel formants were produced in that square. Vowels in black squares were never produced.

citation form map contains hyperarticulated vowels that lie largely on the periphery of vowel space, and the casual speech map has many reduced and centralized vowels. This is a consequence both of reduction inherent to conversational speech and reduced non-stressed syllables, as well as schwa vowels present in highly frequent function words.

When compensating for altered auditory feedback, a subject might either rely on citation form vowels (because he is in a "citation form" mode of speech) or he might access all vowels in his repertoire. Both vowel spaces are specific to the individual. The following analyses measure the ability of either vowel space to predict individual performance on this task.

Analyses

Before reviewing the results, consider the plight of a subject in an altered feedback experiment. He is asked to say 'head', but hears himself saying a vowel that is halfway toward the mean of /æ/ space. Let us assume that his speech motor control system has taken in his auditory and somatosensory feedback along with his expectation for that feedback and detected a mismatch. How does he correct it?

This is equivalent to asking, given that the subject has produced formants in a particular bin in vowel space and heard shifted formants in a different bin, where should he move next? Because we know from Chapter 3 that his speech is autocorrelated, as a first approximation, this analysis assumes that he can only move to an adjacent bin.

The speaker chooses the next bin by balancing the following four forces, derived from the goals justified above. The four properties of the subject's background space thought to be good candidates for determining the subject's performance in these 5 experiments were extracted from the data.

1. Vowel density.
   Determine the F1 and F2 that the subject produced, and look up which bin contains these formants. Count the number of vowel tokens in this bin. Dysfluencies were excluded, but all other instances of the vowel were permitted. Multiple vowels from a single word that happened to fall in the same bin were counted as separate vowels.

2. Probability of hearing intended vowel.

$$\frac{\text{\# vowels in } \textit{heard} \text{ bin identified as intended V}}{\text{total \# vowels in } \textit{heard} \text{ bin}}$$

   Determine the F1 and F2 that the subject heard after the formant shift, and look up which bin contains these formants. Find the proportion of vowels in this bin that were identified as the intended vowel.

3. Probability of feeling intended vowel.

$$\frac{\text{\# vowels identified as intended V in } \textit{produced} \text{ bin}}{\text{total \# vowels in } \textit{produced} \text{ bin}}$$

   Determine the F1 and F2 that the subject produced, and look up which bin contains these formants. Find the proportion of vowels in this bin that were identified as the intended vowel.

4. Completeness of compensation.

$$\sqrt{V_{said} - V_{expected}}$$

   Subtract the feedback shift from the current vowel formants to determine the vowel expected in the case of direct and complete compensation. Find the distance between the center of the *produced* bin and the expected bin.

These scores are calibrated so that vowels in bins with *low* scores are easier to produce or reflect a better response to the experiment.

Example.

A brief example should make these calculations clearer.

Suppose that the subject produced the word 'yes' with /ɛ/ F1=405Hz (4.06 Bark) and F2=1672Hz (11.8 Bark).

Splitting the vowel space created from mock interview data into 30 columns in F2 and 30 rows in F1, this set of formants falls in bin (22, 11) – the 22nd column out of 30 in F2 and the 11th row out of 30 in F1. This bin was occupied 27 times during the interview; the vowel density of this bin is 27.

Nine of the tokens in this bin were identified as /ɛ/, the intended vowel. It was also identified as /ɪ/ 12 times, /ɚ/ 2 times, /eɪ/ 2 times, and /ə/ 2 times.

The probability of feeling oneself say the intended vowel is therefore:

$$\frac{\text{\# vowels identified as intended V in } \textit{produced} \text{ bin}}{\text{total \# vowels in } \textit{produced} \text{ bin}} = \frac{9}{27}.$$

Now suppose that the formant shift at this trial was F1 +150Hz, meaning that the subject heard F1=555Hz (5.39 Bark) and F2=1672Hz (11.8 Bark).

The *heard* vowel fell in bin (22,18). This bin was occupied 6 times during the interview, 1 of which was identified as /ɛ/. (Four were identified as /æ/ and one was identified as /ʌ/).

The probability of hearing the intended vowel is therefore:

$$\frac{\text{\# vowels identified as intended V in } \textit{produced} \text{ bin}}{\text{total \# vowels in } \textit{produced} \text{ bin}} = \frac{1}{6}.$$

Finally, suppose that the subject naturally produces an /ɛ/ with a center of F1=505Hz (4.96 Bark) and F2=1720Hz (12.0 Bark). *Complete compensation* is the formant shift added to the vowel center, or

F1 = (505 - 150) = 355Hz; F2 = 1720Hz.

The distance between the expected vowel and the actual vowel is:

$$\begin{aligned} &\sqrt{V_{said} - V_{expected}} \\ =&\sqrt{(355-405)^2 + (1720-1672)^2)} \\ =&\sqrt{2500+2304} = 69.3\text{Hz} \end{aligned}$$

In principle, each of these four properties could independently predict a subject's performance on this task. In reality, some of these quantities turn out to be correlated across experiments, as shown in Tables 6.1 and 6.2. Generally, these four factors are more independent of each other in the citation form background data.

The probabilities of producing and hearing the intended vowel are correlated in part because feedback is altered for only part of the experiment. The produced and heard vowels turn out to be the same, or very similar, for about 50 of the 200 experimental trials. One of these factors should therefore be removed from the analysis.

	density	**completeness**	**Pr(hear intended V)**	**Pr(produce intended V)**
density	1	0.167	-0.007	-0.095
completeness		1	0.538	0.137
Pr(hear intended V)			1	0.289
Pr(produce intended V)				1

Table 6.1: Correlations between properties of experiment and background vowel space from mock interview.

	density	**completeness**	**Pr(hear intended V)**	**Pr(produce intended V)**
density	1	-0.053	0.045	-0.387
completeness		1	-0.027	0.032
Pr(hear intended V)			1	0.580
Pr(produce intended V)				1

Table 6.2: Correlations between properties of experiment and background vowel space from citation form words.

There are two reasons that removing the probability of hearing the intended vowel is a better choice. First, measurement of vowel production is more accurate than measurement of vowel perception. Errors or noise in re-synthesis can lead to poor LPC formant measurement, especially when vowel bounds are determined automatically. Although the accuracy of both types of measurements is reasonable, there are more missing heard vowel measurements than produced vowel measurements. Second, the subject's production is better understood than the subject's perception. The mock interview established production vowel category boundaries, but no direct measurements were made of the subject's perceptual category boundaries. For this reason, production category information is more accurate than perceptual category information. Because collinearity in predictors may lead to instability and inaccuracy when estimating regression coefficients, the probability of hearing the intended vowel was removed from the latter two analyses.

It is not clear why completeness of compensation is correlated with the probability of hearing (and producing) the intended vowel, and it was not omitted from the analyses.

Analysis 1: Pure optimization

There are several methods of modeling the effect of the experiment and background on the subject's path through vowel space. Perhaps the most intuitive way to think of a subject's formant trajectory from trial to trial is as a cost optimization problem. Consider the subject's vowel space from his mock interview, shown in Figure 6.3 above. On the first trial, the subject produces a vowel in one of those bins. Which bin should he occupy on the next trial? The answer is calculated from the four formulas mentioned above.

Starting in the bin associated with the current set of formants, a score associated with each of the four factors was calculated for the current bin and each of the eight adjacent bins using the method just outlined.

For example, consider a bin on the edge of this subject's /ɛ/ region. This bin was occupied by 15 vowels during the casual speech interview. Fourteen of those vowels were /ɛ/ and one was /æ/. During the current trial of a hypothetical experiment shifting the vowel /ɛ/, the subject produces a vowel in this bin, having produced a vowel in the bin immediately to its left during the previous trial. During this trial, the formant shift is 100Hz, and the bin containing the subject's current vowel is 50Hz away from the baseline /ɛ/ vowel. The bin containing the vowel that the subject *hears* was occupied by 5 vowels during the casual speech interview. All were /ɛ/. Scores for the current bin would be calculated as described above:

> The density of this bin is 15.
>
> The distance between this vowel and the "expected" vowel is the distance from the current vowel, in Bark, to the sum of the baseline vowel formants and the formant shift, converted to Bark.
>
> The probability of producing the intended vowel is $\frac{14}{15}$, or 0.93.
>
> The probability of hearing the intended vowel is $\frac{5}{5}$, or 1.

This procedure is repeated for each of the eight bins adjacent to the bin actually occupied during this trial. To normalize, each score was divided by the sum of the scores for that factor across the nine bins (the current bin + the 8 adjacent bins), and the resulting scores were summed across factors, as shown in Equation 6.1.

$$s_b = \sum_f \left(\frac{s_{b,f}}{\sum_b s_{b,f}} \right), \text{where } b = \text{bin, and } f = \text{factor} \tag{6.1}$$

After scores were calculated for each of the 9 bins, the bin with the lowest score was selected as the optimal bin to visit on the next trial. Performance is measured by checking whether the predicted bin was actually visited on the next trial. Iterating this process on every trial across the five experiments, it is easy to calculate how often the subject chooses the optimal bin.

On average, formant production tends to change less than the width of two bins between trials. A histogram of the F1 and F2 differences from trial to trial is shown in Figure 6.4.

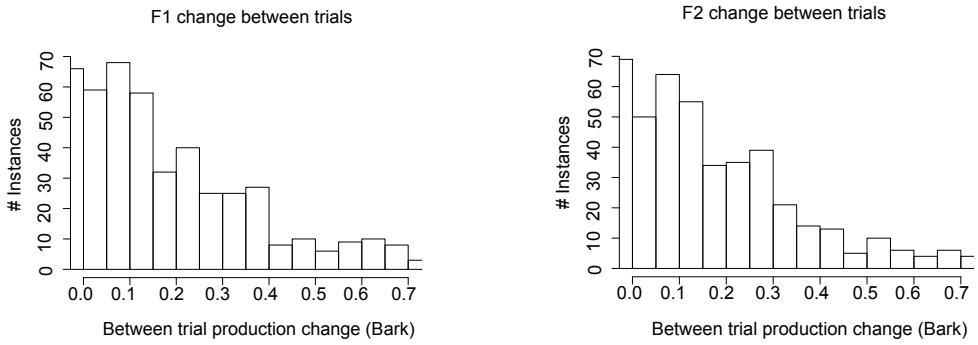


Figure 6.4: Histogram of trial-to-trial change in formants for subject, pooled across 5 experiments. Height of bars indicates number of trials with the indicated amount of formant change.

Because bins are 0.12 Barks wide (F2) and 0.15 Barks tall (F1), the histogram indicates that the subject changes his formant production by a single bin (0.12 to 0.24 Bark in F2, and 0.15 to 0.30 Bark in F1) for many trials. Looking at only the subset of trials for which a subject moves by one bin is a rudimentary way of identifying effects of nearby formants on the actual choice of formants.

Results. Calculated this way, a pure optimization analysis is not successful at predicting the bin that the subject will occupy next. The correct bin is chosen 9.9% of the time. Given that there are 9 possible bins, this performance is no better than chance (11%).

Of course, this analysis relied on a number of simplifications that may explain its failure in prediction. First, and most critically, all four predictors were weighted equally. If the predictors were not of equal import, this assumption could lead to poor prediction. For example, if the density of the surrounding bins were relatively unimportant, but hearing oneself say the intended vowel were very important, then predictions in high density areas, where vowel density improves bins' scores irrespective of the likelihood that the bin contains good representatives of the intended vowel, would be very poor. Second, formant scores are crude: every vowel in a bin has the same set of scores on every measure but completeness. Only the completeness of com-

pensation score does not depend on vowel density and can be calculated for a point rather than a region in vowel space. Third, the subject is assumed to change his vowel formants by, at most, a single bin between trials. It is clear from the histogram that this is not a very accurate simplication.

The second analysis addresses some of these issues.

Analysis 2: Linear regression

The previous analysis predicted the direction of compensation in the next vowel based on the formant shift and the current vowel's background density, completeness of compensation, and the probability of producing the intended vowel in the 8 adjacent bins in vowel space. Weighing these four predictors evenly, a cost was computed for each surrounding bin, and the lowest-cost bin was predicted. This model's predictive power was very poor. It may have been poor either because these properties of the subject's background vowel space do not predict his compensation, or because the analysis was crude. The quality of prediction may be at fault because it was based only on the four evenly-weighted predictors, even though there was no particular reason for weighing them evenly, or because this subject does not usually move by only a single bin between trials.

As an aside, one would anticipate that these predictors will have different weights based on the results of Preliminary Experiment 2. In that experiment, a speaker could compensate fully for small shifts in auditory feedback without producing or hearing a sound that is confusable with an adjacent vowel. But as the size of the formant shift increased, complete compensation required producing a sound that *was* confusable with an adjacent vowel. The partial compensation found in this study shows that most speakers will accept hearing themselves producing an adjacent vowel before they will accept saying an adjacent vowel. This might be a product of the *weighting* of these four quantities.

This linear regression model sketched below allows predictors to be weighted by different amounts. However, highly correlated predictors can lead to inconsistencies in predictor weights, and several of the predictors are correlated. To deal with this problem, the probability of hearing the intended vowel (i.e., the probability that the shifted vowel sounded like the intended vowel) was removed from the analysis. Both completeness of compensation and the probability of hearing the intended vowel were left in the analysis, as they capture two different aspects of the background vowel space.

In order to find the relative weights of the three remaining predictors, this analysis builds a linear regression model that uses the previous set of formants plus a linear combination of the three predictor scores at the current bin to predict the next set of formants.

Including the previous set of formants as a regressor is a decision that needs to be considered carefully. If any of the other predictors (completeness, density,

etc.) predict the previous formant, then much of the variance that might have been attributed to those predictors may instead be attributed to the previous formant. This means that the significance of the three background predictors will be less than it would have been had the previous formant not been included. However, in this analysis, it seems appropriate to include the previous formant, because any variables that remain significant once it is included are predictive of the current formant and *not* the previous formant.

The model is set up with the following equation:

$$\begin{aligned} F_{i+1} &= \beta_1 \cdot density \\ &+ \beta_2 \cdot Pr(\text{completeness}) \\ &+ \beta_3 \cdot Pr(\text{feel intended V}) \\ &+ \beta_4 \cdot F_i \\ &+ \beta_5 \cdot \exp \cdot density \\ &+ \beta_6 \cdot \exp * completeness \\ &+ \beta_7 \cdot \exp \cdot Pr(\text{feel intended V})) \end{aligned}$$

All predictors were normalized.

Results. As in the other experiments, this analysis is designed to predict only a single parameter, and for this reason, F1 and F2 are analyzed separately using R's `glm` function. The subject's joint formant trajectory over the course of the experiment is also of interest, and is considered in Analysis 3.

For F1 across experiments, there were main effects of vowel density ($\beta = -0.27$, $Pr(>|\ t\ |) < 0.0001$), F1 at the previous trial ($\beta = 0.27$, $Pr(>|\ t\ |) < 0.0001$), completeness of compensation ($\beta = 0.14$, $Pr(>|\ t\ |) = 0.03$), the probability of producing the intended vowel ($\beta = 0.14$, $Pr(>|\ t\ |) = 0.0001$), and experiment. A number of two-way interactions were significant as well: between vowel density and the /ɛ/ to /æ/, /ɛ/ -250Hz and /u/ experiments, between completeness of compensation and the /u/ experiment, and producing the intended vowel and experiment type (/ɛ/ to /æ/, /ɛ/ -250Hz, and /ʌ/ -250Hz).

For F2 across experiments, there were marginal effects of density ($\beta = -0.088$, $Pr(>|\ t\ |) = 0.052$) and the probability of producing the intended vowel ($\beta = -0.09$, $Pr(>|\ t\ |) < 0.096$), and main effects of 4 out of 5 experiments (/ɛ/ to /ɪ/: $\beta = 0.73, Pr(>|\ t\ |) < 0.0001$; /ɛ/ -250Hz: $\beta = 0.18, Pr(>|\ t\ |) = 0.0005$, /ʌ/ -250Hz: $\beta = -1.11, Pr(>|\ t\ |) < 0.0001$; /u/ +300Hz: $\beta = -0.30, Pr(>|\ t\ |) < 0.0001$). There were also two-way interactions between vowel density and the /ʌ/ experiment, between completeness of compensation and the /u/ experiment, and between the probability of producing the intended vowel and the /ɛ/ to /æ/, /ɛ/ -250Hz and the /u/ +300Hz experiments. All coefficients from the regression analysis are listed in Tables 6.3 and 6.4.

This analysis suggests that vowel density, at least, may be used by subjects to determine where to go on the next trial. However, there is enough dependence on

F1, casual speech	**Estimate**	**Std. Error**	**t value**	$Pr(>\|t\|)$
density	-0.27	0.0327	-8.16	< 1.5e-15 ***
completeness	0.14	0.064	2.2	0.027 *
Pr(produce intended V)	0.14	0.037	3.90	0.0001 ***
previous F1	0.25	0.038	6.64	5.95e-11 ***
ɛ to ɪ exp	-0.21	0.040	-5.2	2.48e-07 ***
ɛ to æ exp	-0.25	0.29	-0.88	0.38
ɛ -250Hz exp	0.41	0.05	8.78	<2e-16 ***
ʌ -250Hz exp	0.45	0.08	5.73	1.5e-08 ***
u +300Hz exp	-1.44	0.091	-15.9	< 2e-16 ***
ɛ to æ exp · density	0.18	0.052	3.4	0.0006 ***
ɛ -250 exp · density	-0.11	0.049	-2.3	0.022 *
ʌ -250 exp · density	0.05	0.062	0.88	0.38
u +300 exp · density	0.43	0.048	8.98	< 2e-16 ***
ɛ to æ exp ·completeness	0.42	0.46	0.90	0.36
ɛ -250 exp · completeness	-0.018	0.08	-0.22	0.83
ʌ -250 exp · completeness	0.082	0.13	0.61	0.54
u +300 exp · completeness	-0.37	0.085	-4.32	1.74e-05 ***
ɛ to æ exp · Pr(produce V)	-0.19	0.07	-2.5	0.0102 *
ɛ -250 exp · Pr(produce V)	-0.14	0.045	-3.07	0.0022 **
ʌ -250 exp · Pr(produce V)	-0.36	0.067	-5.4	8.55e-08 ***
u +300 exp · Pr(produce V)	-0.34	0.050	-0.683	0.49

Table 6.3: Predictors in linear regression predicting F1 in casual speech.

experiment type that it is important to confirm these results by performing separate regressions for each experiment. Results from those regressions are summarized in Tables 6.5 and 6.6, and listed separately in Appendix B.

F2, casual speech	**Estimate**	**Std. Error**	**t value**	$Pr(>\|t\|)$
density	-0.088	0.045	-1.94	0.052 .
completeness	0.13	0.089	1.41	0.16
Pr(produce intended V)	-0.09	0.051	-1.67	0.096 .
previous F2	0.013	0.033	3.93	9.19e-05 ***
ɛ to ɪ exp	0.73	0.061	12.0	< 2e-16 ***
ɛ to æ exp	0.60	0.41	1.48	0.14
ɛ -250 exp	0.18	0.052	3.49	0.0005 ***
ʌ -250 exp	-1.11	0.11	-10.0	< 2e-16 ***
u +300 exp	-0.30	0.065	-4.66	3.71e-06 ***
ɛ to æ exp · density	-0.083	0.072	-1.15	0.25
ɛ -250 exp · density	-0.014	0.069	-0.20	0.84
ʌ -250 exp · density	0.31	0.086	3.63	0.0003 ***
u +300 exp · density	0.15	0.066	2.26	0.024 *
ɛ to æ exp ·completeness	-0.55	0.65	-0.85	0.40
ɛ -250 exp · completeness	-0.11	0.11	-0.99	0.323
ʌ -250 exp · completeness	-0.39	0.19	-2.08	0.038
u +300 exp · completeness	-0.13	0.12	-1.09	0.277 **
ɛ to æ exp · Pr(produce V)	0.29	0.10	2.8	0.0057 **
ɛ -250 exp · Pr(produce V)	0.19	0.063	3.01	0.00027 ***
ʌ -250 exp · Pr(produce V)	-0.07	0.093	-0.71	0.476
u +300 exp · Pr(produce V)	-0.41	0.071	-5.77	1.2e-08 ***

Table 6.4: Predictors in linear regression predicting F2 in casual speech.

F1 results	**ɛ to ɪ**	**ɛ to æ**	**ɛ -250Hz**	**ʌ -250Hz**	**u +300Hz**
density	***	***	***	NO	NO
completeness	***	NO	NO	NO	*
Pr(produce intended V)	***	***	*	***	NO
previous F1	***	***	***	***	***

Table 6.5: Summary of regressor coefficients in model predicting F1 in casual speech from a mock interview. Factors in boldface were significant as main effects in the pooled experiment analysis. Cells marked with * are significant with Pr(>| t |) between 0.01 and 0.05. Cells marked with ** are significant with Pr(>| t |) between 0.001 and 0.01. Cells marked with *** are significant with Pr(>| t |) <0.001.

F2 results	**ɛ to ɪ**	**ɛ to æ**	**ɛ -250Hz**	**ʌ -250Hz**	**u +300Hz**
density	NO	NO	***	**	NO
completeness	NO	NO	NO	NO	NO
Pr(produce intended V)	NO	*	***	NO	**
previous F2	***	***	***	***	NO

Table 6.6: Summary of regressor coefficients in model predicting F2 in casual speech from a mock interview. Factors in boldface were significant as main effects in the pooled experiment analysis. Cells marked with * are significant with Pr(>| t |) between 0.01 and 0.05. Cells marked with ** are significant with Pr(>| t |) between 0.001 and 0.01. Cells marked with *** are significant with Pr(>| t |) <0.001.

Summary tables 6.5 and 6.6 show the contributions of the four background factors. Previous F1 is a good predictor of the current F1, above and beyond the contributions of the other three factors. This finding was not surprising given the small between-trial distances traveled between trials, illustrated in the first histogram of Figure 6.4. Similarly, the previous F2 was a significant predictor of the current F2 in four out of the five experiments. In addition, actually saying the intended vowel was important to F1 prediction in four out of the five experiments, but was less important to F2 prediction, appearing as significant in only 3 out of 5 experiments. This fits with general intuitions about F1 and F2: F1 is closely tied to jaw height, whereas the relation between F2 and articulator position is more complicated. As for the other predictors, results for vowel density were mixed: density was a significant predictor of F1 in 3 out of the 5 experiments, and a significant predictor of F2 in 2 out of the 5 experiments. It is interesting that casual speech vowel density, a property that varies from individual to individual, predicts the current F1 but not the current F2. Perhaps surprisingly, completeness of compensation was even less helpful than vowel density. It was a significant predictor of F1 in only two out of the five experiments, and it was never a significant predictor of F2. Apparently, even this single subject does not consistently try to oppose the shift in formant feedback. However, it must be acknowledged that this subject produced small compensation responses, meaning that his formant trajectory may be noisier than that of a subject who opposed the formant shift more consistently.

For the experiment that altered only F1, density and producing the intended vowel were important, as was the relationship to the previous F1. But the two experiments that altered F2 behaved quite differently. In the /ʌ/ experiment, F2 was predicted by density and the previous F2, while in the /u/ experiment, F2 was predicted by the probability of hearing the intended vowel. No predictor was significant for both F2 experiments.

This is a potentially promising set of results. Perhaps when compensating, it is not enough to produce a vowel that one has produced before. That vowel may also have to be a member of the correct vowel category. Certainly this result is supported by the phoneme experiment in Chapter 5, which showed that compensation was influenced by speakers' phonological inventories. It is also encouraging that there seems to be a cross-experimental pattern in the probability of producing the intended vowel: it seems to be quite important in predicting F1, and sometimes important in predicting F2.

Although predictors in this analysis were permitted to have different weights, and subjects were not constrained to move by any particular amount between trials, there were a number of major problems with this regression. First and foremost, the significance of the previous F1 or F2 shows that there may be autocorrelation between trials, especially during initial no-shift trials and ramp trials. A linear regression of this sort is not designed to handle time series data. Second, because the subject was permitted to change his formant production by any amount between trials, it was not

possible to model the subject's performance as an optimization problem. Third, F1 and F2 trajectories are not independent and should not be analyzed separately. The third analysis addresses all of these shortcomings.

Analysis 3: Optimization of regression coefficients

The second analysis, which used a linear combination of the four predictive properties of a subject's background space to predict the subject's formants at the next trial from his formants at the current trial, yielded some interesting insights into his performance. In particular, it found that the four properties of the subject's phoneme background were not equally important to the prediction of the following set of formants. Producing the intended vowel in F1 seemed to matter a lot, while vowel density and completeness of compensation were not consistently important. These findings suggest that one reason for the inaccuracy of prediction in the optimization analysis was that all predictors in that analysis were weighted equally. However, a deficiency of the second analysis lies in the interaction with experiment type. When the five experiments were combined and coefficients were co-estimated, there was a significant interaction of experiment type with the four predictors. That is, the predictors had different weights, and those weights varied by experiment; in some experiments, frequency had a larger coefficient than the probability of producing the intended vowel, and in other experiments the reverse was true. A number of possible conclusions can be drawn from such inconsistencies.

1. The subject is weighing different predictors by different amounts, depending on the experiment.
2. The subject is not using any of these predictors systematically.
3. These predictors are actually controlled by a different set of variables that were not measured.
4. The subject compensates for F1 and F2 jointly.
5. The subject often chooses his next formant not because it is the best formant *overall* based on frequency or hearing the right vowel, but rather because it is the best formant among the nearby options.

Based on Analysis 2, (1) seems to be true, and suggests that the subject's weights are task-dependent. This is not unreasonable: it implies that compensation depends on the formant being altered, the word being altered, and the amount of alteration. If (2) is true, then there are no patterns to find in this subject's data; his path is random. If (3) is true, then the amount of variance unaccounted for in the regression model would be small. (4) and (5) are interesting possibilities that have not been ruled out because the regression model deals with compensation in F1 and F2 independently,

and does not treat compensation as optimization. Instead of viewing compensation as a process of taking stock of the available formant choices and choosing the best of those options, the regression model notes the conditions in the bins actually chosen, and tries to find a pattern based on those choices.

It is therefore worth trying to combine the optimization and regression approaches, and attempt to find optimal weightings for coefficients while still taking the subject's other options into account at every trial.

To do this, the third analysis optimizes the coefficients of the three relevant predictors based on the transition probability between the previous trial and the current trial. Importantly, this analysis takes into account *all* other positions in vowel space, not just the immediately adjacent bins. This can be thought of as a linear regression based on the *transitional probability* between the previous bin and the current bin in vowel space. Because this method operates on the 2-D vowel space grid, it does not have to deal with one formant at a time like the second regression analysis did.

The model is a generalized linear regression, but because the data depends on both position and time, it is more complex than a standard model that can be evaluated with R's `glm` function. There are three key pieces: X, the predictors, β, the coefficients of those predictors, and D, the observed formant trajectory. It is set up as follows.

D, the observed formant trajectory, is converted to a sequence of bins in a 30x30 grid overlaid on the subject's vowel space.

X is the set of matrices containing predictor values at each trial, taking into account the formant shift. For each trial, there is a 30×30 matrix of bins in vowel space. Given the position in vowel space at that trial and the formant shift at that trial, it is straightforward to evaluate each of the three predictors for each bin, giving a 30(rows) $\times$ 30(columns) $\times$ 3(predictors) matrix for each trial. For ease of notation, equations refer to the first predictor, density, as X_1; the second predictor, completeness, as X_2; and the third predictor, the probability of producing the intended vowel, as X_3. Missing trials were ignored.

β is the vector of weights corresponding to the three predictors.

Π is the product of β and X. If the relative weights for each of the predictors were known, it would be possible to calculate the probability of occupying any particular bin during a trial using the equation below. The weights are not known, of course. Solving for them is the task of R's `optim` function, as described below.

Π corresponds to a set of transition probability matrices. Each matrix is 30x30, with each element corresponding to a bin in vowel space. In order to make the value of each bin a valid probability, the product of β and X is first transformed to a quantity η greater than 0 by taking its exponent. The resulting positive number is then divided by the sum of all other η across bins (in the current trial). This is a standard method of transforming from real numbers to probabilities. The transition probability p_t of each bin is calculated with equation 6.2.

$$p_t(r,c) = \frac{\exp\left(-\sum_j \beta_j X_{j,r,c}\right)}{\sum_r \sum_c \exp\left(-\sum_j \beta_j X_{j,r,c}\right)} \tag{6.2}$$

where r=row, c=column. Recall that X_j is density when j=1; X_j is completeness of compensation when j=2; and X_j is the probability of producing the intended vowel when j=3.

Once the transitional probability matrices Π are constructed, it is easy to calculate the likelihood of the formant trajectory across trials. The likelihood L of the observed formants is the product of the transitional probabilities across trials. Computationally, this is implemented using the sum of the log of the transitional probabilities, as shown in equation 6.3.

$$L = \sum_t (\log(p_t)) \tag{6.3}$$

The `optim` function in R finds the β coefficients that maximize the log of this cross-trial likelihood. It uses the Nelder Mead simplex algorithm to converge on the optimal coefficients. This algorithm works by building a 3-dimensional triangle and evaluating the equation at each of the triangle vertices; the first set of vertices are pre-selected. At each iteration, the vertex of the triangle with the worst (highest) score is replaced with a new guess. In this way, the triangle stretches and shrinks its way from the start position to the function minimum. As with all optimization algorithms, it is possible to get "stuck" in a local minimum. To avoid this, the `optim` function was run with several sets of parameter start values, including (0,0,0). It converged to the same parameter values, no matter where it started, suggesting that local minima were not a problem.

Results. Because the experiments were not performed consecutively, each experiment needed its own transition probability matrix. Before pooling the data across experiments, optimal coefficients for the three predictors were found for each experiment individually. The results are laid out in Table 6.7.

Table 6.7 shows that there are some general patterns in coefficients across experiments in spite of considerable variability. Vowel density and completeness of compensation correlated negatively with the observed formant trajectory, while the probability of hearing the intended vowel correlated positively with the observed sequence of vowel formants. The /ʌ/ experiment is the only experiment whose coefficients are not consistent with this general trend.

The negative coefficients on the completeness predictor imply that subjects had a tendency to produce vowel formants that are a *small* distance from the formants expected in the case of complete compensation. The positive coefficients on the

Predictor	ɛ to ɪ	ɛ to æ	ɛ -250Hz	ʌ -250Hz	u +300Hz
density	-1.07	-0.309	-0.137	0.067	-0.015
completeness	-1.27	-3.93	-0.189	-1.03	-0.520
Pr(produce intended V)	1.98	4.74	0.667	-0.944	3.00

Table 6.7: β coefficients of vowel density, completeness of compensation, and the probability of hearing the intended vowel that are associated with the maximum likelihood of the formant trajectory data. Data from each experiment was analyzed separately. Density information was extracted from casual speech.

Pr(produce intended V) predictor imply that subjects tended to produce formants with a *high* probability of being identified as the intended vowel. Both of these results are expected. However, the negative coefficients on the density predictor imply that subjects aim to produce formants in *low* density areas of vowel space. This is strange, and calls into question the use of this background vowel density. A different vowel density, derived from the subject's citation form vowels, will be used in the next section and compared to these results.

Although the signs on the predictors were fairly uniform across experiments, the exact values of β that best solve this equation are highly dependent on the experiment.

Predictor	β, pooled data
density	0.045
completeness	-1.21
Pr(produce intended V)	1.47

Table 6.8: Results of the optimization analysis, data pooled across all 5 experiments. β coefficients of vowel density, completeness of compensation, and the probability of hearing the intended vowel that are associated with the maximum likelihood of the formant trajectory data. Density information was extracted from casual speech.

Surprisingly, Table 6.8 shows that when data are pooled across experiments, the sign of the coefficient in front of density flips: instead of being negative, it becomes weakly positive. This new coefficient suggests a very small overall effect of vowel density, in which subjects slightly prefer producing vowels in high-density regions.

The discrepancy between the results in the individual experiments and the pooled data suggests that the pooled data might be unduly influenced by one or two experiments that are out of line with the others. To verify the stability of these coefficients, the pooled data optimization was re-run with one experiment excluded. The excluded experiment rotated from the first to the fifth, so that the effect of each individual experiment on the overall coefficients could be observed. The results of this cross-validation are summarized in Table 6.9.

Predictor	-ɛ to ɪ	-ɛ to æ	-ɛ -250Hz	-ʌ -250Hz	-u +300Hz
density	0.047	0.041	0.060	-0.90	0.049
completeness	-1.12	-0.856	-1.54	-1.15	-1.29
Pr(say intended V)	1.19	0.80	1.71	2.08	1.45

Table 6.9: Results of the optimization analysis using data from all experiments *except* the one indicated in the column heading. β coefficients of vowel density, completeness of compensation, and the probability of hearing the intended vowel that are associated with the maximum likelihood of the formant trajectory data are listed. Density information was extracted from casual speech.

Table 6.9 shows that the density coefficient is stable, except in the /ʌ/ experiment, and that the completeness and Pr(produce intended V) coefficients are not especially affected by any experiment. These latter two coefficients are also generally less stable when one experiment is removed. Again, the informativeness of the background vowel space is very low. The next section re-evaluates these three analyses with the background vowel space extracted from the 360 citation form vowels that this subject produced during his control trials.

Citation form vowels

On the whole, this subject's conversational vowel space turns out not to be a good predictor of his trial-to-trial compensation. It does very poorly at prediction in the pure optimization analysis, and finds task-dependent coefficients in the regression analysis. It is possible that the casual speech vowel space is not a good predictor because it is unrelated to his compensation; perhaps compensation is actually determined by factors such as personality, mood, perceptual ability, or physiology. Before dismissing the notion that properties of a subject's background vowel space help to predict variation in compensation between trials and sessions, consider one additional possibility, that the casual speech vowel space is not an appropriate background. From the perspective of exemplar models of speech production, vowels from citation form words are much more likely to be influenced by other citation vowels than by vowels tagged as casual speech. In this view, when the subject produces laboratory speech with isolated words, he enters "laboratory speech mode" and accesses citation form vowels more readily than casual speech vowels.

With this in mind, this section investigates a modified hypothesis, that a subject's compensation depends on the subset of his background vowel space used in citation form words, laboratory speech, or clear speech. As a proxy for citation form speech density, 360 vowels from CVC monosyllables, produced during this subject's control condition, are used, as shown in Figure 6.3.

Optimization Results.

Using the same procedure as in Analysis 1 above and the citation form vowel space, and considering only those trials where the subject changed his vowel production by one bin or less, the pure optimization analysis chooses the correct bin in vowel space in 7.7% of trials. Again, this is no better than chance. This analysis does not predict the subject's formant trajectory over trials more effectively than the analysis using the casual speech vowel space for prediction.

Regression Results.

In the regression predicting F1 from the citation vowel background space, there was a main effect of the previous F1 ($\beta = 0.39, Pr(>|\ t\ |) < 2e-16$). There were also main effects for all experiments except the linguistic /ɛ/ to /æ/ experiment. There was also a significant interaction between vowel density and the /ɛ/ -250Hz experiment.

In the regression predicting F2 from the citation vowel background space, there was a main effect of the previous F2 ($\beta = 0.087, Pr(>|\ t\ |) = 0.016$), and of all experiments except the linguistic /ɛ/ to /æ/ experiment. There was also a marginally significant interaction between vowel density and the /u/ +300Hz experiment. This result suggests that there are substantial differences in behavior from experiment to experiment, justifying the need for experiment-level analyses. Details of the cross-experimental regression analysis are listed in Tables 6.10 and 6.11.

When the individual experiments are analyzed separately, we find that different predictors are significant in different experiments, as indicated in Tables B.1 to B.10 in Appendix B. A summary of these tables, presented in Tables 6.12 and 6.13, show that the only main effects significant in (almost) all experiments are previous F1 and previous F2. In F2, completeness of compensation is a significant predictor of the next formant in 3 out of 5 experiments.

Citation speech, F1	**Estimate**	**Std. Error**	**t value**	$Pr(>\|t\|)$
density	0.052	0.038	1.33	0.183
completeness	0.067	0.071	0.94	0.346
Pr(produce intended V)	0.023	0.045	0.50	0.62
previous F1	0.39	0.043	9.23	< 2e-16 ***
ɛ to ɪ exp	-0.14	0.044	-3.27	0.0011 **
ɛ to æ	-0.15	0.33	-0.46	0.65
ɛ -250Hz exp	0.48	0.051	9.48	< 2e-16 ***
ʌ -250Hz exp	0.37	0.09	4.13	4.11e-05 ***
u +300Hz exp	-1.10	0.22	-4.91	< 1.11e-06 ***
ɛ to æ· density	-0.001	0.049	-0.026	0.98
ɛ -250Hz · density	-0.18	0.058	-3.18	0.00152 **
ʌ -250Hz · density	0.014	0.29	-0.048	0.96
u +300Hz · density	0.060	0.18	0.33	0.74
ɛ to æ· completeness	0.15	0.55	0.27	0.79
ɛ -250Hz · completeness	0.005	0.094	0.055	0.956
ʌ -250Hz · completeness	-0.0027	0.16	-0.017	0.986
u +300Hz · completeness	-0.11	0.29	-0.37	0.72
ɛ to æ· Pr(produce V)	-0.058	0.061	-0.95	0.345
ɛ -250Hz · Pr(produce V)	-0.068	0.068	-1.00	0.315
ʌ -250Hz · Pr(produce V)	0.025	1.23	0.204	0.838
u +300Hz · Pr(produce V)	-0.074	0.35	-0.21	0.83

Table 6.10: Predictors in linear regression predicting F1 in citation form speech.

Citation speech, F2	**Estimate**	**Std. Error**	**t value**	$Pr(>\|t\|)$
density	-0.060	0.052	-1.15	0.25
completeness	0.047	0.09	0.50	0.61
Pr(produce intended V)	0.011	0.061	0.179	0.858
previous F2	0.087	0.036	2.41	0.016 *
ɛ to ɪ exp	0.74	0.06	11.4	< 2e-16 ***
ɛ to æ	0.35	0.45	0.79	0.43
ɛ -250Hz exp	0.25	0.052	4.92	1.07e-06 ***
ʌ -250Hz exp	-1.05	0.12	-8.76	< 2e-16 ***
u +300Hz exp	-0.57	0.29	-1.92	0.055 .
ɛ to æ· density	0.034	0.066	0.51	0.612
ɛ -250Hz · density	0.11	0.079	1.41	0.16
ʌ -250Hz · density	-0.24	0.39	-0.61	0.542
u +300Hz · density	-0.43	0.25	-1.75	0.081 .
ɛ to æ· completeness	0.11	0.74	0.15	0.88
ɛ -250Hz · completeness	-0.097	0.13	-0.77	0.44
ʌ -250Hz · completeness	-0.29	0.21	-1.35	0.178
u +300Hz · completeness	-0.34	0.39	-0.88	0.38
ɛ to æ· Pr(produce V)	0.050	0.083	0.61	0.54
ɛ -250Hz · Pr(produce V)	0.14	0.091	1.51	0.132
ʌ -250Hz · Pr(produce V)	0.008	0.16	0.05	0.959
u +300Hz · Pr(produce V)	0.167	0.47	0.36	0.722

Table 6.11: Predictors in linear regression predicting F2 in citation form speech.

F1 results	**ɛ to ɪ**	**ɛ to æ**	**ɛ -250Hz**	**ʌ -250Hz**	**u +300Hz**
density	NO	***	**	NO	NO
completeness	NO	NO	NO	NO	NO
Pr(produce intended V)	NO	*	NO	**	NO
previous F1	***	***	***	***	***

Table 6.12: Summary of regressor coefficients in model predicting F1 in citation speech. Factors in boldface were significant as main effects in the pooled experiment analysis. Cells marked with * are significant with Pr(>| t |) between 0.01 and 0.05. Cells marked with ** are significant with Pr(>| t |) between 0.001 and 0.01. Cells marked with *** are significant with Pr(>| t |) <0.001.

F2 results	**ɛ to ɪ**	**ɛ to æ**	**ɛ -250Hz**	**ʌ -250Hz**	**u +300Hz**
density	NO	NO	NO	NO	NO
completeness	**	***	**	NO	NO
Pr(produce intended V)	NO	**	**	NO	NO
previous F2	***	***	***	***	NO

Table 6.13: Summary of regressor coefficients in model predicting F2 in citation speech. Factors in boldface were significant as main effects in the pooled experiment analysis. Cells marked with * are significant with Pr(>| t |) between 0.01 and 0.05. Cells marked with ** are significant with Pr(>| t |) between 0.001 and 0.01. Cells marked with *** are significant with Pr(>| t |) <0.001.

Analysis 3: Optimization of regression coefficients

The third analysis, which optimized predictor weights based on trial and subject-specific information, was repeated using the citation vowels as a background space. Table 6.14 shows that coefficients using this background space were far more stable across experiments.

Predictor	**ɛ to ɪ**	**ɛ to æ**	**ɛ -250Hz**	**ʌ -250Hz**	**u +300Hz**
density	0.322	0.384	0.197	0.191	-0.799
completeness	-1.67	-2.66	-0.442	-1.32	-2.37
Pr(produce intended V)	1.77	1.13	1.44	1.64	4.18

Table 6.14: Predictors in optimization analysis finding the weights associated with the maximum likelihood of the data and subject-specific information about vowel density, completeness of compensation, and the probability of hearing the intended vowel.

Table 6.14 shows that there is considerable uniformity in coefficients across trials. Vowel density and the probability of hearing the intended vowel correlated positively with the observed formant trajectory, while completeness of compensation correlated negatively with the observed sequence of vowel formants. The /u/ experiment is the only experiment whose coefficients are not consistent with this general trend.

As in the interview analysis, using this background, this subject still prefers to produce vowels in areas with *high* vowel density where he is *likely* to feel himself saying the intended vowel. He also prefers vowels that are a *small* distance from the formants he would have produced had he compensated completely. At maximum shift, these quantities are different from each other, as shown by the low correlation between them (0.17 for casual vowels; -0.05 for citation vowels).

Discussion

The hypothesis underpinning this chapter was that individual variation in a talker's response to formant alteration depends on (1) the subject's starting formant values; (2) the feedback shift; and (3) the subject's native vowel space. This chapter outlined a series of exploratory analyses undertaken to determine whether this hypothesis is reasonable. Two of these analyses evaluate F1 and F2 jointly, and one evaluates F1 and F2 separately. At this point, it is not clear which model is more appropriate for analyzing the data. Both yield interesting patterns.

The first analysis viewed the compensation problem as an optimization of competing costs, based on experimental conditions at each trial and the subject's vowel density map. This analysis did not find any of these factors to be important. When they were weighted equally, this method was no better than chance at predicting how the subject's formants changed from trial to trial.

The second analysis phrased compensation as a linear regression problem. In that model, the talker's next formant was predicted from a linear combination of the current formant, the experimental conditions at each trial, and the talker's background vowel space. The analysis solved for the optimal weighting of the three factors for F1 and F2 separately. It showed that not only are some factors more important than others, but the identity of the important factors *changes* from experiment to experiment. This partly explains why the initial optimization analysis failed. However, this regression model included the previous formant as a regressor, meaning that it required the current trial to be independent of all other trials, given the previous trial. This choice reduced the significance of the remaining predictors, perhaps accounting for their poor predictive power. Some variance in the background predictors that was correlated with the previous formant may have been attributed to the previous formant rather than the individual predictors. Consequently, any predictors that were significant in Analysis 2 were significant beyond the contribution of the previous formant. Nonetheless, because this data set potentially contains autocorrelation, missing variables, and collinear variables, further analyses are warranted. One regression analysis should not be the final word on individual compensation patterns.

To take advantage of the cost-minimizing framing of the compensation problem from Analysis 1 and the differential regressor weighting from Analysis 2, Analysis 3 minimizes a cost function based on scores associated with all possible "next moves" in the compensation experiment. The scores are based on a linear regression, allowing weighting of individual predictors to be optimized. It also was able to evaluate a joint function of F1 and F2. According to this analysis, the subject consistently preferred to be closer to the formants he would have produced had he compensated completely, and consistently produced vowels that were likely to feel like the intended vowel. Vowel density did not much affect this subject's formant trajectory when calculated from his casual speech, but he did seem to aim for areas of high vowel density in his citation form speech. Even in this analysis, there was some variation from experiment to experiment.

Although some general patterns in Analysis 3 were clear, the coefficients of the three predictors varied in magnitude between experiments. It is possible that this variability reflects the influence of other, unmeasured variables. This chapter has explored only a few dimensions out of many along with individuals actually differ, including personality traits and sociolinguistic affiliation. Certain personality traits are likely to influence one's willingness to attend to one's own voice. For example, individuals differ in their ability to self-monitor (Snyder, 1974), and autism tends to be correlated with hyperacute hearing (Bonnel et al., 2003). In addition, sociolinguistic affiliation may make certain vowels particularly salient: midwesterners might be especially inclined to correct for hearing themselves producing a fronted /u/, which they would associate with sounding 'Californian'. Work is in progress to investigate the influence of social and personality variables on compensation.

Apart from these sociolinguistic factors, other variables related to physiology,

perceptual ability, and learning biases were not measured. The subject's perceptual boundaries across vowel space were inferred from his production, but a more complete analysis would use a vowel continuum categorization task to generate a map of vowel labels for this subject. In addition, differences in the subject's vocal tract shape surely constrain the subject's vowel space and are likely to affect learning and vowel targets, but vocal tract shape was measured only indirectly in these analyses. Similarly, hearing acuity, which was not measured here, has been shown to affect compensation (Perkell, 2007). Finally, neuroanatomical differences have been shown to affect learning of new sounds (Wong et al., 2008), and so there is good reason to suppose that the volume or topography of structures in the subject's brain could affect compensation behavior. It is therefore possible that subtle physical, psychological, or neuroanatomical differences affect the mental comparison of observed with expected feedback. Because background vowel space turns out to be only one factor controlling an individual's compensation for altered auditory feedback, it is worth exploring the connection between all of these factors and speech motor control.

In spite of these caveats, Analysis 3 demonstrated that vowel density from this subject's *citation form* vowel formants is a better predictor of his formant trajectory than his *casual speech* vowel formants. This suggests that when this subject evaluates which formants to produce next, he first considers formants that he has produced before *in a similar speaking situation*. Even in this low-level task, the subject appears to have indexical knowledge of his intended speech style. This finding provides support for exemplar models of speech processing.

Chapter 7

General Discussion

The experiments outlined in this dissertation had two goals. One was to investigate why altered auditory feedback is incomplete. In particular, I asked whether certain aspects of a talker's linguistic system affect low-level monitoring of learned sounds and learning of new sounds. The four experiments described here did so by altering auditory feedback in real time, causing talkers to hear the vowels in their own voices with shifted formants.

Measuring differences in compensation for altered auditory feedback can help us understand several phenomena of interest in linguistic research. First, differences in compensation may be linked to natural variation in speech production. Variation in production is thought to provide the substrate for diachronic phonological changes. In addition, these results can help us understand how vowel targets change with development. Finally, they can help to unify models of speech processing in two domains: psycholinguistic models of speech processing, which deal with high-level, word-by-word message planning, and speech motor control models, which deal with low-level execution and monitoring of speech plans. These two types of models are not incompatible *per se*, but until now there was not enough information to determine how they should be unified. Results of the experiments performed in this dissertation are a step toward collecting that information.

The preliminary experiments in Chapter 3 showed that (1) the Feedback Alteration Device used in this dissertation works similarly to other feedback alteration devices in the literature; (2) confirmed that compensation for altered auditory feedback is *incomplete*; and (3) established that altered auditory feedback is less complete for large feedback shifts than for small ones. Part of this incompleteness may simply be a feature of the controller: it may be that speech targets do not change quickly, even in the presence of incongruous feedback. In the arm motor control case, compensation is incomplete, at about 85%, possibly for this reason (Welch, 1971). But compensation in the speech motor control system is even less complete, generally around 25-50%, and that additional incompleteness does deserve an explanation.

I argued that this incompleteness may be interesting from the point of view of

psycholinguistic models of speech production and for models of speech motor control because incomplete compensation may reflect the influence of (a) the lexical inventory; (b) the phonological inventory; (c) articulatory familiarity on the comparison between observed and expected feedback. Understanding the role of these aspects of a talker's linguistic background expands our knowledge of speech motor control systems, and provides information about the common currency used in preparation of a message and the execution of that message, allowing us to merge high-level and low-level models of speech production.

Chapter 4 sought to find out if a talker's lexicon influenced compensation for altered auditory feedback. That experiment measured compensation in altered feedback during production of two nonwords, [tɛg] and [dɛg]. The experiment was designed so that complete compensation for *teg* required speakers to say the word *tag*, whereas complete compensation for *deg* required speakers to say the nonword [dæg]. This experiment failed to find a consistent difference between compensation for *deg* and compensation for *teg*, even when the experimental design encouraged subjects to activate their lexicons. While it remains possible that there is some effect of lexicon on compensation for altered auditory feedback, any such effect would have to be subtle enough to escape detection in this experiment.

Chapter 5 investigated compensation for auditory feedback altered in different parts of the vowel space. Because different vowels are associated with different auditory *and* somatosensory targets, this chapter was able to investigate the role of changes in both target types on compensation for altered auditory feedback. In this experiment, subjects produced monosyllabic CVC words with V = {ɛ, ʌ, or u}. From a phonological perspective, there is only 1 vowel lying within a 2 Bark radius of /u/, while there are 3 vowels lying within 2 Bark of /ɛ/ and 5 vowels lying within 2 Bark of /ʌ/. If subjects were to avoid compensation in areas where they were likely to hear or produce competing vowels, compensation would be less complete for vowels in denser regions than for vowels in sparser regions. In addition, these three vowels differ in somatosensory information. Whereas [u] is articulated with both palatal contact and lip rounding, [ɛ] is articulated with some palatal contact but no lip rounding, and [ʌ] is articulated with little palatal contact and no lip rounding. If subjects were to avoid compensation for vowels with strong somatosensory feedback (because compensating requires accepting one's production of that vowel with unusual somatosensory feedback), compensation would be less complete in vowels with salient somatosensory feedback.

Results of this experiment show that compensation is large for the vowel /u/, and small for both /ɛ/ and /ʌ/. This pattern of compensation magnitudes is consistent with an influence of phonological inventory on compensation: compensation was greatest for the vowel with the fewest neighbors (/u/, with 3 neighbors), smaller for /ɛ/, with 4 neighbors, and smallest for /ʌ/, with 5 neighbors. It is not consistent with somatosensory salience (in which feedback should be least for /u/ followed by /ɛ/ and then /ʌ/). Based on this evidence, I argued that there is some interaction

between a talker's phonological inventory and his speech motor control system.

Other explanations for this pattern cannot be ruled out on the basis of this evidence. Because the experimental designs differed, the three vowels were shifted by three different amounts, though due to their positions in vowel space, these amounts were far more different in Hertz (400Hz, 300Hz, 250Hz) than they were in Bark (0.90Bark, 0.53Bark, 0.50Bark).

A second experiment presented in this chapter showed that subjects compensate for very small (30Hz and 90Hz) feedback shifts in their vowel spaces. Because subjects managed to compensate for feedback changes in which vowels they heard fell within their vowel target regions, vowel expectations must be generated from specific motor commands rather than the syllable in general.

Chapter 6 investigated the contribution of articulatory and acoustic familiarity to compensation using three exploratory analyses. In this case study, I collected a subject's *background vowel space* in two ways. First, the subject produced 360 CVC words with V={i, ɪ, ɛ, æ, a, ʌ, o, u}. Second, formants from 4654 vowels were collected by asking the subject to speak casually during a half-hour mock interview. After the background vowel spaces were collected, the subject participated in 5 altered feedback experiments. This study was limited in that it investigated the behavior of a single participant.

This study used a linear regression analysis to determine whether the formants that the subject produced during each trial of each experiment could be predicted by the formant shift and characteristics of the vowel background. It found that this subject generally tried to avoid producing ambiguous vowels and aimed to produce frequently-articulated vowels in his citation form vowel space. His vowel production was not well correlated with frequency of production in his casual speech vowel space. These results were clear evidence that individual characteristics of a talker will influence the way he will respond to an altered feedback experiment. For example, a subject who happened to produce a vowel near an infrequently used portion of vowel space during Trial 1 may compensate less than he would have had he started in the middle of his vowel region, and a subject whose /ɛ/ and /æ/ were very close together should respond less to an /ɛ/ shift than a subject whose /ɛ/ and /æ/ were well-separated.

There are two major limitations to this study. One is that a perceptual space was not collected. Without it, it is difficult to make strong claims about the importance of producing sounds within a vowel category. Perceptual data was not collected because it would have been labor intensive to ask the subject to categorize vowels produced at every (F1, F2) in vowel space, but a good follow-up study would collect this data. A second limitation is that articulatory data is not collected: perhaps a subject's first concern is not to move the articulators too much. Because the relationship between articulation and acoustics is nonlinear, subjects may be compensating more or less steadily than they appear to be from their vowel formants alone.

Conclusions

The experiments outlined in this dissertation expand our knowledge of speech production and speech monitoring. They also suggest ways of incorporating high-level and low-level models.

Chapter 5 expanded our knowledge of speech motor control in two ways. The correlation between compensation and phonological neighborhood density demonstrated that the speech motor control system is influenced by the phonological system during adjustment of articulatory plans. Within-vowel-category compensation demonstrated that expectations are generated from specific motor commands rather than syllable-level articulatory plans. To accommodate the first result, the phonological inventory has been added to the diagram of the speech motor control system. It appears as one of the inputs to the abstract speech plan. The second result confirms that an internal model is likely to be part of the speech motor control system, with motor commands as input and an expectation as output.

Chapter 6 demonstrated that speech targets are specific to individual talkers. The results of the regression analysis, which showed that subjects tend to produce vowels in dense areas of citation form vowel space and unambiguous vowel formants, can expand on models of speech motor control. It seems that a talker's expectation is fed not only by acoustic and somatosensory information generated by the internal model, but also by higher-level knowledge about the intended phoneme (or syllable), and contextual knowledge about the speech register. Knowledge of speech register is especially surprising at this low level, suggesting that even self-monitoring of auditory and somatosensory feedback is language and experience-specific.

I argue that high-level and low-level models of speech production are linked as in Figure 7.1. Some parts of this unified model are still vague, opening multiple new routes for experimentation.

One open question is whether the two types of models share phonemic, syllabic, or gestural representations. These three studies do not provide enough information to decide among these choices. The second is whether processes performed during low-level speech monitoring feed back up to high-level planning. It would be interesting to look for frequency effects resulting from altered auditory feedback; category boundary changes have already been noted in multiple studies (e.g., Shiller et al., 2009; Ostry, Darainy, Mattar, Wong, & Gribble, 2010). Another avenue would be to incorporate models of phonological short term memory, which have their own views on the "chunks" that are processed during speech perception and production (Jacquemot & Scott, 2006).

Some high-level models of speech production already incorporate some form of self-monitoring in their models (Levelt et al., 1999). But the picture of self monitoring suggested by the Levelt et al. model is too simplistic, and too high-level, to account for the altered feedback data observed here. They suggest that speakers attend to their speech on the sound level and the phonemic level, and make adjustments at the

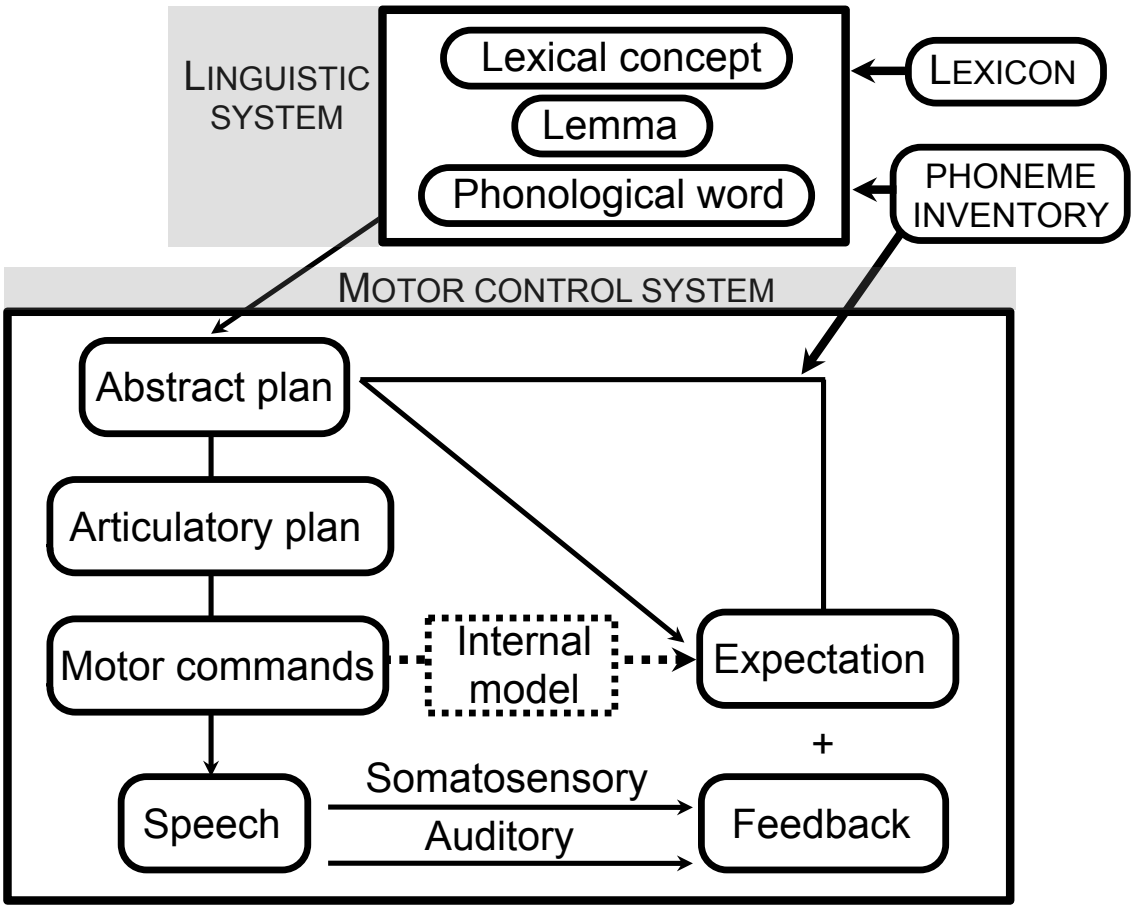


Figure 7.1: Unified model of speech production.

conceptual level when one of them has gone awry, reselecting the intended message, word, etc. While self-monitoring surely does happen on this level, there is additional self-monitoring that is not accounted for in this model. The model could be updated to reflect this new information.

I hope that this framework can be used to understand accommodation, the process by which interlocutors begin to sound more like each other. A growing body of work explores the social factors that promote accommodation and what phonetic properties become more similar as talkers accommodate (Pardo, 2006; Babel, 2009). The cognitive mechanism responsible for accommodation has not been researched, but it is likely similar to the mechanism responsible for compensation for altered auditory feedback. In both cases, talkers change their articulation of speech sounds on account of incoming speech. There is one important and interesting difference between them: in adaptation, talkers *oppose* changes to the speech they hear, and in accommodation, talkers *assimilate* toward the speech they hear. The difference al-

most certainly arises from processing of self-produced versus other-produced speech. In compensation, subjects are reacting to self-initiated speech, and they have access to the motor commands that generated the speech, allowing them to generate an accurate expectation for what they are going to hear. In accommodation, listeners receive an incoming acoustic signal, but not the associated motor commands. Because incoming speech affects a listener's outgoing speech, there must be some connection between them. The nature of this connection remains open to future research.

Perhaps relatedly, researchers involved in pitch perturbation have thought extensively about differences in processing between self-produced speech and other-produced speech. Although some subjects in pitch perturbation experiments partially oppose manipulations of auditory feedback, some subjects actually *follow* the direction of the manipulation, producing a higher pitch when they hear the pitch of their voices raised. The tendency to follow manipulations is greater for greater pitch shifts. One reason subjects might follow a pitch perturbation is that, for large deviations from normal, subjects no longer believe that the voice they are hearing is their own. If this turns self-produced speech into other-produced speech, then they are more likely to *accommodate* toward the voice they hear rather than oppose it and reach their own pitch target.

Caveats

The experiments outlined here concluded that several high-level aspects of talkers' native languages contributed to incomplete compensation for altered auditory feedback. Still, neither phonological nor lexical nor articulatory effects can fully explain why compensation is so poor; there is still much "missing" compensation to account for. Two good candidates for explaining partial compensation have not been fully investigated. One candidate is social factors; certain people seem to be better compensators than others. It is likely that attitudes about one's dialect, tendency to self-monitor, or degree of extroversion might weigh heavily in any particular talker's decision to compensate (or not).

A second factor is the *speed* of change in feedback. Altered visual feedback experiments show that subjects reach a compensation asymptote, beyond which their perception does not change, when their visual field is suddenly tilted by 30 degrees. Subjects will continue to compensate past that asymptote, however, if the 30 degree tilt is introduced at the slower rate of 1.4 degrees per minute (Ebenholtz & Callan, 1980). In the speech domain, perhaps different speakers have different methods of making large changes to acoustics and articulation, but have similar methods of making smaller changes. If it is the nature of controllers to be stable – after all, it would be undesirable to have one's vowel targets shifting wildly every time one spoke on a bad phone connection – compensation might be more complete (and slowly adapting) if input were changed very slowly. The best way of testing this hypothesis is beyond the range of current technology; one would have to wear a portable formant shifting

device that altered incoming formants incrementally over the course of several days. Perhaps under these conditions learning is more complete and more permanent.

Future work

All of the experiments described in this dissertation document how young, male, largely undergraduate California English speakers with normal hearing and speaking skills react to altered auditory feedback. This has been a useful exercise; there is enough variability even within this uniform population to make some interesting inferences about the workings of the speech motor control system. Specifically, the experiments in this dissertation suggest that compensation is likely insensitive to lexical inventory, but is sensitive to phonological inventory and the way an individual uses his or her vowel space.

A logical next step is to verify that these trends hold across other populations of different ages and language backgrounds. A good test case would be a language with an identical lexical inventory but a different phonological inventory, particularly in the [ɪ]-[ɛ]-[æ] region. A good language for testing this hypothesis is New Zealand English, which has [ɪ] and [ɛ] but not [æ]. Investigating these vowels may be especially interesting because some of these vowels are in flux, and speakers of different ages may represent them differently. Comparing compensation and generalization of adaptation to surrounding vowels would shed light on how vowel representations in the two dialects of English differ. An experiment to compare compensation in this population to compensation in speakers of American English speakers is currently being planned.

This framework also holds promise for learning about the nature of language disorders. For example, populations with apraxia of speech have damaged either their internal models or the abstract plans that feed into their internal models. Even when apraxic speakers know they want to produce the word 'cat', they cannot remember how to move their articulators to form the appropriate sounds. In severe cases, they flail about, trying to initiate the /k/ in multiple ways without managing to produce a closure. Another good example is populations with aphasia, who have experienced a breakdown in this speech production system following a stroke or brain injury. Aphasia leaves people with language-specific difficulties, from word selection to articulation to speech monitoring. Of interest to models of speech motor control are *fluent* persons with aphasia[1], who seem to have a disconnection between speech perception and production. These persons speak without effort, but have a host of surprising production and comprehension difficulties. Many persons with Wernicke's aphasia produce nonsense words and, crucially, do not realize that they are doing so. The fact that they produce nonsense words suggests a high-level lexical or syllable selection problem, and their inability to realize that they are producing jargon suggests

[1]Fluent aphasias are usually typed as Wernicke's or Conduction aphasia.

that they have an inability to self-monitor, at least on the word level[2]. Such talkers should not be able to compensate for altered auditory feedback, and their responses should be entirely uncorrelated with their feedback shift. Others with aphasia have comprehension problems or difficulty in repeating sentences that they have heard, even though their speech production is fluent when self-initiated. Learning about how these subjects respond to altered auditory feedback – for example, whether they can compensate for changes in pitch or loudness without being able to compensate for formant shifts, or whether they produce frequent vowels that are unconnected to the experimental condition – can help speech pathologists and neuropsychologists learn more about how the aphasic brain processes language. This experiment is currently underway.

[2]Some research indicates that fluent aphasics are able to discriminate phonemes, however.

References

Abbs, J. H., & Gracco, V. L. (1984). Control of complex motor gestures: orofacial muscle responses to load perturbations of lip during speech. *Journal of Neurophysiology*, *51*(4), 705-723.

Ash, S. (2006). The North American Midland as a dialect area. In T. E. Murray & B. L. Simon (Eds.), *Language variation and change in the American Midland: a new look at 'Heartland' English* (p. 33-). John Benjamins.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412.

Babel, M. (2009). *Phonetic and social selectivity in speech accommodation*. Unpublished doctoral dissertation, University of California at Berkeley, Berkeley, CA.

Baddeley, A. D. (1992). Working memory. *Science*, *255*(5044), 556-559.

Beckman, M. E., Jung, T.-P., Lee, S., de Jong, K., Krishnamurthy, A. K., Ahalt, S. C., et al. (1995). Variability in the production of quantal vowels revisited. *Journal of the Acoustical Society of America*, *97*(1), 471-490.

Beddor, P. S., & Strange, W. (1982). Cross-language study of perception of the oral–nasal distinction. *Journal of the Acoustical Society of America*, *71*(6), 1551-1561.

Bell-Berti, F., Raphael, L. J., Pisoni, D. B., & Sawusch, J. R. (1979). Some relationships between speech production and perception. *Phonetica*, *36*, 373-383.

Berg, T., & Abd-El-Jawad, H. (1996). The unfolding of suprasegmental representations: a cross-linguistic perspective. *Journal of Linguistics*, *32*, 291-324.

Bhushan, N., & Shadmehr, R. (1999). Computational nature of human adaptive control during learning of reaching movements in force fields. *Biological Cybernetics*, *81*(1), 39-60.

Blakemore, S.-J., Goodbody, S. J., & Wolpert, D. M. (1998). Predicting the consequences of our own actions: the role of sensorimotor context estimation. *The Journal of Neuroscience*, *18*(18), 7511-7518.

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355-387.

Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). Neural representations

and mechanisms for the performance of simple speech sequences. *Journal of Cognitive Neuroscience*, *22*(7), 1504-1529.

Bonnel, A., Mottron, L., Peretz, I., Trudel, M., Gallun, E., & Bonnel, A.-M. (2003). Enhanced pitch sensitivity in individuals with autism: A signal detection analysis. *Journal of Cognitive Neuroscience*, *15*(2), 226-235.

Bradlow, A. R. (2002). Confluent talker- and listener-oriented forces in clear speech production. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology 7* (p. 241-274). Mouton de Gruyter.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*(4), 2299-2310.

Brand, M., Rey, A., & Peereman, R. (2003). Where is the syllable priming effect in visual word recognition? *Journal of Memory and Language*, *48*, 435-443.

Broad, D. J. (1976). Toward defining acoustic phonetic equivalence for vowels. *Phonetica*, *33*(401), 424.

Browman, C., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*, 155-180.

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

Burnett, T. A., & Larson, C. R. (2002). Early pitch-shift response is active in both steady and dynamic voice pitch control. *Journal of the Acoustical Society of America*, *112*(3), 1058-1063.

Carré, R., Ainsworth, W. A., Jospa, P., Maeda, S., & Pasdeloup, V. (2001). Perception of vowel-to-vowel transitions with different formant trajectories. *Phonetica*, *58*(3), 163-178.

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

Chen, T.-M., Dell, G. S., & Chen, J.-Y. (2004). A cross-linguistic study of phonological units: Syllables emerge from the statistics of Mandarin Chinese, but not from the statistics of English. *Cogsci2004*.

Chiba, T., & Kajiyama, M. (1941). *The vowel, its nature and structure*. Tokyo: Tokyo-Kaiseikan Publishing Company Ltd.

Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of american english. *Journal of the Acoustical Society of America*, *118*(3), 1661-1676.

Cohen, A., Slis, I. H., & 'T Hart, J. (1963). Perceptual tolerances of isolated Dutch vowels. *Phonetica*, *9*, 65-78.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1983). A language specific comprehension strategy. *Nature*, *304*, 159-160.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283-321.

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, *27*, 124-142.

Diehl, R. L. (2000). Searching for an auditory description of vowel categories. *Phonetica*, *57*, 267-274.

Ebenholtz, S. M., & Callan, J. W. (1980). Tilt adaptation as a feedback control process. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 413-432.

Fahey, R. P., & Diehl, R. L. (1996). The missing fundamental in vowel height perception. *Perception and Psychophysics*, *58*, 725-733.

Fahey, R. P., Diehl, R. L., & Traunmüller, H. (1996). Perception of back vowels: Effects of varying f1-f0 distance. *Journal of the Acoustical Society of America*, *99*, 2350-2357.

Fairbanks, G. (1955). Selective vocal effects of delayed auditory feedback. *Journal of Speech and Hearing Disorders*, *20*, 333-346.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton & Company.

Farnetani, E. (1999). Coarticulation and connected speech processes. In W. J. Hardcastle & J. Laver (Eds.), *Handbook of phonetic sciences* (p. 371-404). Malden, MA: Blackwell Publishers Ltd.

Ferrand, L., Seguí, J., & Grainger, J. (1996). Masked priming of word and picture naming: the role of syllabic units. *Journal of Memory and Language*, *35*, 708-723.

Ferrand, L., Segui, J., & Humphreys, G. W. (1997). The syllable's role in word naming. *Memory and Cognition*, *25*(458-470).

Flemming, E., & Johnson, S. (2007). Rosa's roses: Reduced vowels in american english. *Journal of the International Phonetic Association*, *37*, 83–96.

Fowler, C. A. (1994). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception and Psychophysics*, *55*(6), 597-610.

Fowler, C. A., & Brancazio, L. (2000). Coarticulation resistance of American English consonants and effects on transconsonantal vowel-to-vowel coarticulation. *Language and Speech*, *43*, 1-41.

Fowler, C. A., & Saltzman, E. L. (1993). Coordination and coarticulation in speech production. *Language and Speech*, *36*(2,3), 171-195.

Fowler, C. A., & Turvey, M. T. (1981). Immediate compensation in bite-block speech. *Phonetica*, *37*(5-6), 306-326.

Fox, R. A. (1982). Individual variation in the perception of vowels: implications for a perception-production link. *Phonetica*, *39*, 1-22.

Fujimura, O., & Kakita, Y. (1979). Remarks on the quantitative description of

lingual articulation. In B. Lindblom & S. E. G. Öhman (Eds.), *Frontiers of speech communication research* (p. 17-24). London: Academic Press.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110-125.

Ghosh, S. S., Tourville, J. A., & Guenther, F. H. (2008). A neuroimaging study of premotor lateralization and cerebellar involvement in the production of phonemes and syllables. *Journal of Speech, Language, and Hearing Research*, *51*, 1183-1202.

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, *105*(2), 251-279.

Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use. In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (p. 159-207). Berlin: Mouton de Gruyter.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*, 585-612.

Gordon, A. M., & Soechting, J. F. (1995). Use of tactile afferent information in sequential finger movements. *Experimental Brain Research*, *107*, 281-292.

Gottfried, T. L. (1984). Effects of consonant context on the perception of French vowels. *Journal of Phonetics*, *12*, 91-114.

Gracco, V. L., & Löfqvist, A. (1994). Speech motor coordination and control: Evidence from lip, jaw, and laryngeal movements. *The Journal of Neuroscience*, *14*(11), 6585-6597.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*(3), 594-621.

Guenther, F. H. (2003). Neural control of speech movements. In A. S. Meyer & N. O. Schiller (Eds.), *Phonetics and phonology in language comprehension and production: differences and similarities* (p. 209-240). Berlin: Mouton de Gruyter.

Guenther, F. H., Epsy-Wilson, C. Y., Boyce, S. E., Matthies, M. L., Zandipour, M., & Perkell, J. S. (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *Journal of the Acoustical Society of America*, *105*(5), 2854-2865.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, *96*, 280-301.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, *105*(4), 611-633.

Hagiwara, R. (1997). Dialect variation and formant frequency: The American English

vowels revisited. *Journal of the Acoustical Society of America*, *102*(1), 655-658.

Hall-Lew, L. (2009). *Ethnicity and phonetic variation in a San Francisco neighborhood*. Unpublished doctoral dissertation, Stanford University, Palo Alto, CA.

Hawkins, S., & Stevens, K. N. (1985). Acoustic and perceptual correlates of the non-nasal – nasal distinction for vowels. *Journal of the Acoustical Society of America*, *77*(1560-1575).

Henke, W. L. (1966). *Dynamic articulatory model of speech production using computer simulation*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*(5), 3099-3111.

Hillenbrand, J., & Nearey, T. M. (1999). Identification of resynthesized /hVd/ utterances: effects of formant contour. *Journal of the Acoustical Society of America*, *105*(6), 3509-3523.

Honda, M., & Fujino, A. (2002). Compensatory responses of articulators to unexpected perturbation of the palate shape. *Journal of Phonetics*, *30*, 281–302.

Hose, B., Langner, G., & Scheich, H. (1983). Linear phoneme boundaries for German synthetic two-formant vowels. *Hearing Research*, *9*(1), 13-25.

Houde, J. F. (1997). *Sensorimotor adaptation in speech production*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Houde, J. F., Heinks-Maldonado, T. H., & Nagarajan, S. S. (2006, October 14-18). *Response in auditory cortex predicts subjects' compensations to feedback perturbations during speaking.* (Poster presented at the Society for Neuroscience 2006 meeting (SFN2006))

Houde, J. F., & Jordan, M. I. (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research*, *45*(2), 295-310.

Houde, J. F., & Nagarajan, S. S. (under review). Speech production as state feedback control.

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: An MEG study. *Journal of Cognitive Neuroscience*, *14*(8), 1125-1138.

Jacquemot, C., & Scott, S. K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences*, *10*(11), 480-486.

Johnson, K. (1997). Speech perception without speaker normalization: an exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (p. 145-166). San Diego: Academic Press.

Johnson, K., Flemming, E., & Wright, R. (1993). The Hyperspace Effect: Phonetic targets are hyperarticulated. *Language*, *69*(3), 505-528.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of

talker gender in vowel perception. *Journal of Phonetics*, *27*, 359-384.

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

Jones, J. A., & Munhall, K. G. (2003). Learning to produce speech with an altered vocal tract: The role of auditory feedback. *Journal of the Acoustical Society of America*, *113*(1), 532-543.

Katseff, S., Houde, J. F., & Johnson, K. (2008). Partial compensation in speech adaptation. *Annual Report of the UC Berkeley Phonology Lab*, 444-461.

Keating, P. A., Lindblom, B., Lubker, J., & Kreiman, J. (1994). Variability in jaw height for segments in English and Swedish VCVs. *Journal of Phonetics*, *22*, 407-422.

Kelso, J. A. S., & Tuller, B. (1983). "Compensatory articulation" under conditions of reduced afferent information: A dynamic formulation. *Journal of Speech and Hearing Research*, *26*, 217-224.

Kewley-Port, D. (2001). Vowel formant discrimination II: Effects of stimulus uncertainty, consonantal context, and training. *Journal of the Acoustical Society of America*, *110*(4), 2141-2155.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244-247.

Kubozono, H. (1989). The mora and syllable structure in Japanese: evidence from speech errors. *Language and Speech*, *32*(3), 249-278.

Kubozono, H. (1996). Speech segmentation and phonological structure. In T. Otake & A. Cutler (Eds.), *Phonological structure and language processing: Cross-linguistic studies* (p. 77-94). Berlin: Mouton de Gruyter.

Kuehn, D. P., & Moll, K. (1972). Perceptual effects of forward coarticulation. *Journal of Speech and Hearing Research*, *15*, 654-664.

Kuehn, D. P., Templeton, P. J., & Maynard, J. A. (1990). Muscle spindles in the velopharyngeal musculature of humans. *Journal of Speech and Hearing Research*, *33*, 488-493.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, *50*(2), 93-107.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America*, *100*(4), 2425-2438.

Kureta, Y., Fushimi, T., & Tatsumi, I. F. (2006). The functional unit in phonological encoding: Evidence for moraic representation in native Japanese speakers. *Journal of experimental psychology. Learning, memory, and cognition*, *32*(5), 1102-1119.

Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of North American English: phonetics, phonology, and sound change: a multimedia reference tool*. Mouton

de Gruyter.

Laganaro, M., & Alario, F.-X. (2006). Are syllables represented and retrieved during phonological encoding? An integration of psycholinguistic and neurolinguistic studies. *Journal of Memory and Language*, *55*(178-196).

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, *14*, 677-709.

Larson, C. R., Altman, K. W., Liu, H., & Hain, T. C. (2008). Interactions between auditory and somatosensory feedback for voice F0 control. *Experimental Brain Research*, *187*, 613-621.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1-75.

Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, *50*, 239-269.

Liberman, A. M., Cooper, F. S., Shankweiler, D. S., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431-461.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358-368.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, *35*(11), 1773-1781.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (Vol. 55, p. 403-439). Norwell, MA: Kluwer Academic Publishers.

Lindblom, B., Lubker, J., & Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics*, *7*(147-161).

Lindblom, B., Sussman, H. M., Modarresi, G., & Burlingame, E. (2002). The Trough Effect: Implications for speech motor programming. *Phonetica*, *59*, 245-262.

Liu, D., & Todorov, E. (2007). Evidence for the flexible sensorimotor strategies predicted by optimal feedback control. *The Journal of Neuroscience*, *27*(35), 9354-9368.

Logan, G. D., & Crump, M. J. C. (2010). Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science*, *330*, 683-686.

MacDonald, E. N., Goldberg, R., & Munhall, K. G. (2010). Compensations in response to real-time formant perturbations of different magnitudes. *Journal of the Acoustical Society of America*, *127*(2), 1059-1068.

Maeda, S. (1991). On articulatory and acoustic variability. *Journal of Phonetics*, *19*, 321-331.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29-63.

McAulay, R. J., & Quatieri, T. F. (1991). Low-rate speech coding based on the

sinusoidal model. In S. Furui & M. M. Sondhi (Eds.), *Advances in speech signal processing* (Vol. 76, p. 165-208). New York: Marcel Dekker.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1-86.

McFarland, D. H., & Baum, S. R. (1995). Incomplete compensation to articulatory perturbation. *Journal of the Acoustical Society of America*, *97*(1865-1873).

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227-234.

Morsella, E., & Miozzo, M. (2002). Evidence for a cascade model of lexical access in speech production. *Journal of experimental psychology: Learning, memory, and cognition*, *28*(2), 555-563.

Munhall, K. G., Löfqvist, A., & Kelso, J. A. S. (1994). Lip-larynx coordination in speech: Effects of mechanical perturbations to the lower lip. *Journal of the Acoustical Society of America*, *95*(6), 3605-3616.

Munhall, K. G., MacDonald, E. N., Byrne, S. K., & Johnsrude, I. (2009). Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *Journal of the Acoustical Society of America*, *125*(1), 384-390.

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research*, *47*, 1048-1058.

Nasir, S. M., & Ostry, D. J. (2006). Somatosensory precision in speech production. *Current Biology*, *16*, 1918-1923.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*(5), 2088-2113.

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, *80*(5), 1297-1308.

Nielsen, K. (2008). *The specificity of allophonic variability and its implications for accounts of speech perception.* Unpublished doctoral dissertation, University of California, Los Angeles, Los Angeles, CA.

Niziolek, C. A. (2010). *The role of linguistic contrasts in the auditory feedback control of speech.* Unpublished doctoral dissertation, Massachusetts Institute of Technology and Harvard University, Cambridge, MA.

Niziolek, C. A., & Guenther, F. H. (2009). The influence of perceptual categories on auditory feedback control during speech. *NeuroImage*, *47*(1), S39-S41.

Norris, D. G. (1994). Shortlist: a connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189-234.

Norris, D. G., McQueen, J., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*(3), 299-325.

Öhman, S. E. G. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America*, *41*(2), 310-320.

O'Seaghdha, P., & Chen, J.-Y. (2009). Toward a language-general account of word production: The Proximate Units Principle. In *Proceedings of the 31st Annual Conference of the Cognitive Society* (p. 68-73). Austin, TX: Cognitive Science Society.

Ostry, D. J., Darainy, M., Mattar, A. A. G., Wong, J., & Gribble, P. L. (2010). Somatosensory plasticity and motor learning. *The Journal of Neuroscience*, *30*(15), 5384-5393.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, *119*, 2382-2393.

Perkell, J. S. (2007). Sensory goals and control mechanisms for phonemic articulations. *Proceedings of the International Conference of Phonetic Sciences*, *XVI*, 169-174.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Perrier, P., Vick, J., et al. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics*, *28*, 233-272.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., et al. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *Journal of the Acoustical Society of America*, *116*(4), 2338-2344.

Perkell, J. S., Matthies, M. L., Svirsky, M. A., & Jordan, M. I. (1993). Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: a pilot "motor equivalence" study. *Journal of the Acoustical Society of America*, *93*(5), 2948-2961.

Perkell, J. S., & Nelson, W. L. (1985). Variability in production of the vowels /i/ and /a/. *Journal of the Acoustical Society of America*, *77*(5), 1869-1875.

Perrier, P., Lœvenbruck, H., & Payan, Y. (1996). Control of tongue movements in speech: the equilibrium point hypothesis perspective. *Journal of Phonetics*, *24*(53-75).

Peterson, G. E., & Barney, H. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*, 175-184.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, *28*, 96-103.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, *29*, 434-446.

Pick, Jr., H. L., Siegel, G. M., Fox, P. W., Garber, S. R., & Kearney, J. K. (1989). Inhibiting the Lombard effect. *Journal of the Acoustical Society of America*, *85*(2), 894-900.

Pile, E. J. S., Dajani, H. R., Purcell, D. W., & Munhall, K. G. (2007). Talking under conditions of altered auditory feedback: Does adaptation of one vowel generalize to other vowels? In *Proceedings of the International Conference of Phonetic Sciences* (Vol. XVI, p. 645-648).

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, *13*, 253-260.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics*, *15*(2), 285-290.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., et al. (2007). Buckeye corpus of conversational speech (2nd release). *[www.buckeyecorpus.osu.edu]*, *Columbus, OH*(Department of Psychology), Ohio State University (Distributor).

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*, 146-159.

Pörschmann, C. (2000). Influences of bone conduction and air conduction on the sound of one's own voice. *Acta Acustica united with Acustica*, *86*(6), 1038-1045.

Potter, R. K., & Steinberg, J. C. (1950). Toward the specification of speech. *Journal of the Acoustical Society of America*, *22*, 807-820.

Purcell, D. W., & Munhall, K. G. (2006a). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *Journal of the Acoustical Society of America*, *120*(2), 966-977.

Purcell, D. W., & Munhall, K. G. (2006b). Compensation following real-time manipulation of formants in isolated vowels. *Journal of the Acoustical Society of America*, *119*(4), 2288-2297.

Purcell, D. W., & Munhall, K. G. (2008). Weighting of auditory feedback across the english vowel space. *Proceedings of the 8th International Seminar on Speech Production*.

Quatieri, T. F. (2002). *Discrete-time speech processing: Principles and practice*. Upper Saddle River: Prentice Hall PTR.

Quatieri, T. F., & McAulay, R. J. (1986). Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *ASSP-34*(6), 1449-1464.

Quatieri, T. F., & McAulay, R. J. (1992). Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *40*(3), 497-510.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*, 249-284.

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

Schiller, N. O., Costa, A., & Colomé, A. (2003). Phonological encoding of single words: In search of the lost syllable. In C. Gussenhoven & N. Warner (Eds.),

Papers in laboratory phonology 7 (p. 35-59). Berlin: Mouton de Gruyter.

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M., & Bienkowski, M. (1982). Pre- and postlexical loci of contextual effects on word recognition of contextual effects on word recognition. *Memory and Cognition*, *12*, 315-328.

Shaiman, S., & Gracco, V. L. (2002). Task-specific sensorimotor interactions in speech production. *Experimental Brain Research*, *146*, 411-418.

Shattuck-Hufnagel, S. (1987). The role of word-onset consonants in speech production planning: New evidence from speech error patterns. In E. Keller & M. Gopnik (Eds.), *Motor and sensory processes of language* (p. 17-51). Hillsdale, NJ: Erlbaum.

Shiller, D. M., Sato, M., Gracco, V. L., & Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *Journal of the Acoustical Society of America*, *125*(2), 1103-1113.

Shockey, L. (2003). *Sound patterns of spoken English.* Malden, MA: Blackwell Publishers Ltd.

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology*, *30*(4), 526-537.

Sober, S. J., & Sabes, P. N. (2003). Multisensory integration during motor planning. *The Journal of Neuroscience*, *23*(18), 6982-6992.

Sober, S. J., & Sabes, P. N. (2005). Flexible strategies for sensory integration during motor planning. *Nature Neuroscience*, *8*(4), 490-497.

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, *43*, 482-489.

Stevens, K. N. (1969). Crosslanguage study of vowel perception. *Language and Speech*, *12*, 1-23.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In P. B. Denes & E. E. David (Eds.), *Human communication: A unified view* (p. 51-66). McGraw-Hill.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*, 3-45.

Stone, M., & Lundberg, A. (1996). Three-dimensional tongue surface shapes of English consonants and vowels. *Journal of the Acoustical Society of America*, *99*(6), 3728-3737.

Story, B. H. (2005). A parametric model of the vocal tract area function for vowel and consonant simulation. *Journal of the Acoustical Society of America*, *117*(5), 3231-3254.

Strange, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of the Acoustical Society of America*, *85*(5), 2135-2153.

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, *74*(3), 695-705.

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., & Cooper, F. S. (1970). Motor theory of speech perception revisited: a reply to Lane's critical review. *Psychological Review*, *77*(3), 234-249.

Sussman, H. M., & Westbury, J. R. (1981). The effects of antagonistic gestures on temporal and amplitude parameters of anticipatory labial coarticulation. *Journal of Speech and Hearing Research*, *24*, 16-24.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of american english vowels. *Journal of the Acoustical Society of America*, *79*(4), 1086-1100.

Thomas, E. (2004). Rural Southern white accents. In E. W. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, & C. Upton (Eds.), *A handbook of varieties of English* (Vol. 1, p. 300-324). Mouton de Gruyter.

Tilsen, S. (2009). Effects of syllable stress on articulatory planning: evidence from a stop-signal experiment. *Journal of Phonetics*, *33*, 839-879.

Tourville, J. A., & Guenther, F. H. (2010, March 31). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*.

Tourville, J. A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *NeuroImage*, *39*, 1429-1443.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America*, *69*(5), 1465-1475.

Traunmüller, H. (1988). Paralinguistic variation and invariance in the characteristic frequencies of vowels. *Phonetica*, *45*, 1-29.

Tremblay, S., Houle, G., & Ostry, D. J. (2008). Specificity of speech motor learning. *The Journal of Neuroscience*, *28*(10), 2426-2434.

Tremblay, S., Shiller, D. M., & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature*, *423*, 866-869.

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *Journal of the Acoustical Society of America*, *122*(4), 2306-2319.

Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research*, *50*, 1510-1545.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. *Journal of the Acoustical Society of America*, *117*(1), 338-350.

Vousden, J., Brown, G. D. A., & Harley, T. A. (2000). Oscillator-based control of the serial ordering of phonology in speech production. *Cognitive Psychology*, *41*, 101-175.

Watt, D., Llamas, C., & Johnson, D. E. (2010). Levels of linguistic accommodation across a national border. *Journal of English Linguistics*, *38*, 270-289.

Welch, R. B. (1971). Prism adaptation: The "target-pointing effect" as a function of

exposure trials. *Perception and Psychophysics*, *9*, 102-104.

Welch, R. B. (1986). Adaptation of space perception. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. I). Wiley-Interscience.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49-63.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, *76*(1-15).

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, *11*, 1317-1329.

Wong, P. C. M., Warrier, C. M., Penhune, V. B., Roy, A. K., Sadehh, A., Parrish, T. B., et al. (2008). Volume of left heschl's gyrus and linguistic pitch learning. *Cerebral Cortex*, *18*(4), 828-836.

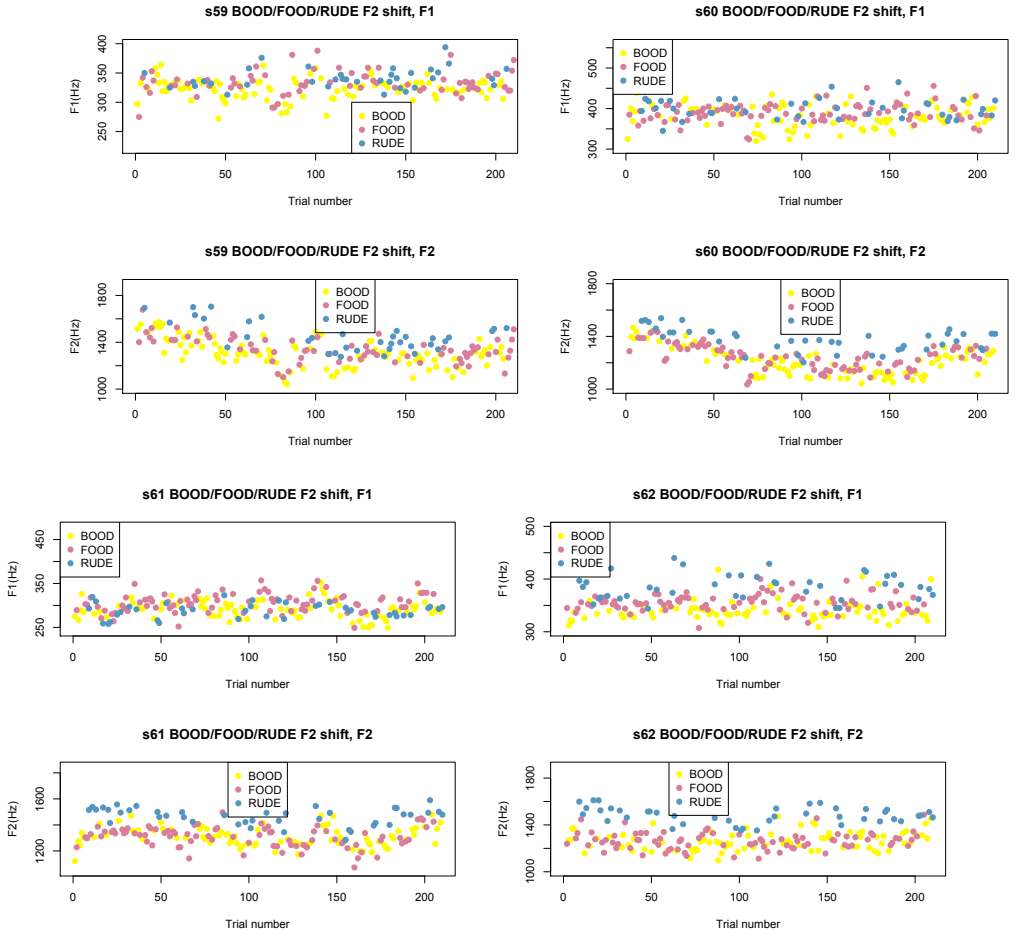
Wright, R. (1997). *Lexical competition and reduction in speech: A preliminary report* (Research on Spoken Language Processing Progress Report No. 21). Bloomington, Indiana: Indiana University.

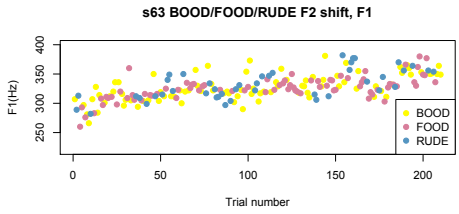
Wright, R. (2004). Factors of lexical competition in vowel articulation. In R. O. J. Local & R. Temple (Eds.), *Papers in Laboratory Phonology VI* (pp. 26–50). Cambridge University Press.

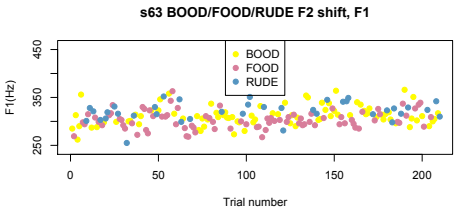
Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, *68*(5), 1523-1525.

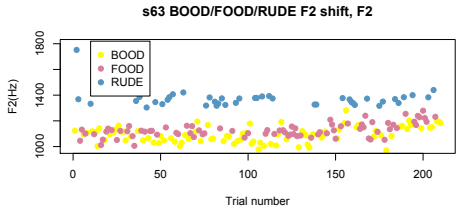
Appendix A

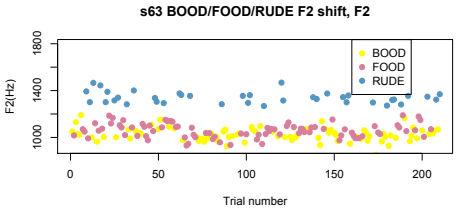
Individual /u/ F2 increase (+300 Hz) results are included below.

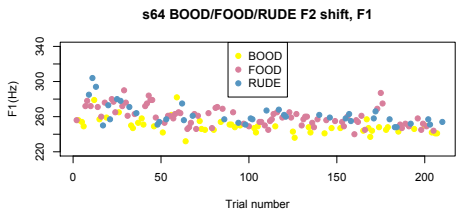


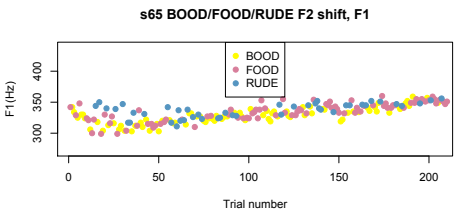

s63 BOOD/FOOD/RUDE F2 shift, F1
F1(Hz)
Trial number
BOOD
FOOD
RUDE

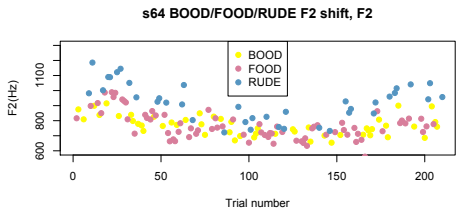

s63 BOOD/FOOD/RUDE F2 shift, F1
F1(Hz)
Trial number
BOOD
FOOD
RUDE

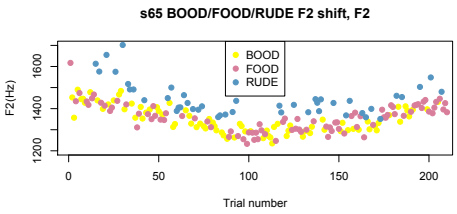

s63 BOOD/FOOD/RUDE F2 shift, F2
F2(Hz)
Trial number
BOOD
FOOD
RUDE

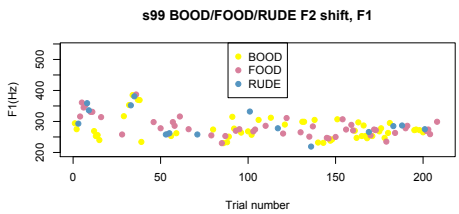

s63 BOOD/FOOD/RUDE F2 shift, F2
F2(Hz)
Trial number
BOOD
FOOD
RUDE

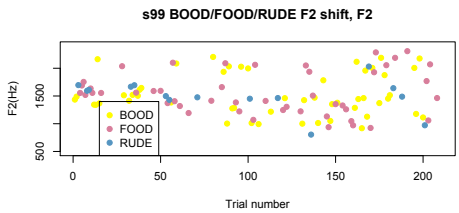

s64 BOOD/FOOD/RUDE F2 shift, F1
F1(Hz)
Trial number
BOOD
FOOD
RUDE


s65 BOOD/FOOD/RUDE F2 shift, F1
F1(Hz)
Trial number
BOOD
FOOD
RUDE


s64 BOOD/FOOD/RUDE F2 shift, F2
F2(Hz)
Trial number
BOOD
FOOD
RUDE


s65 BOOD/FOOD/RUDE F2 shift, F2
F2(Hz)
Trial number
BOOD
FOOD
RUDE


s99 BOOD/FOOD/RUDE F2 shift, F1
F1(Hz)
Trial number
BOOD
FOOD
RUDE


s99 BOOD/FOOD/RUDE F2 shift, F2
F2(Hz)
Trial number
BOOD
FOOD
RUDE

Individual vowel regions for 30Hz and 90Hz ɛ shifts

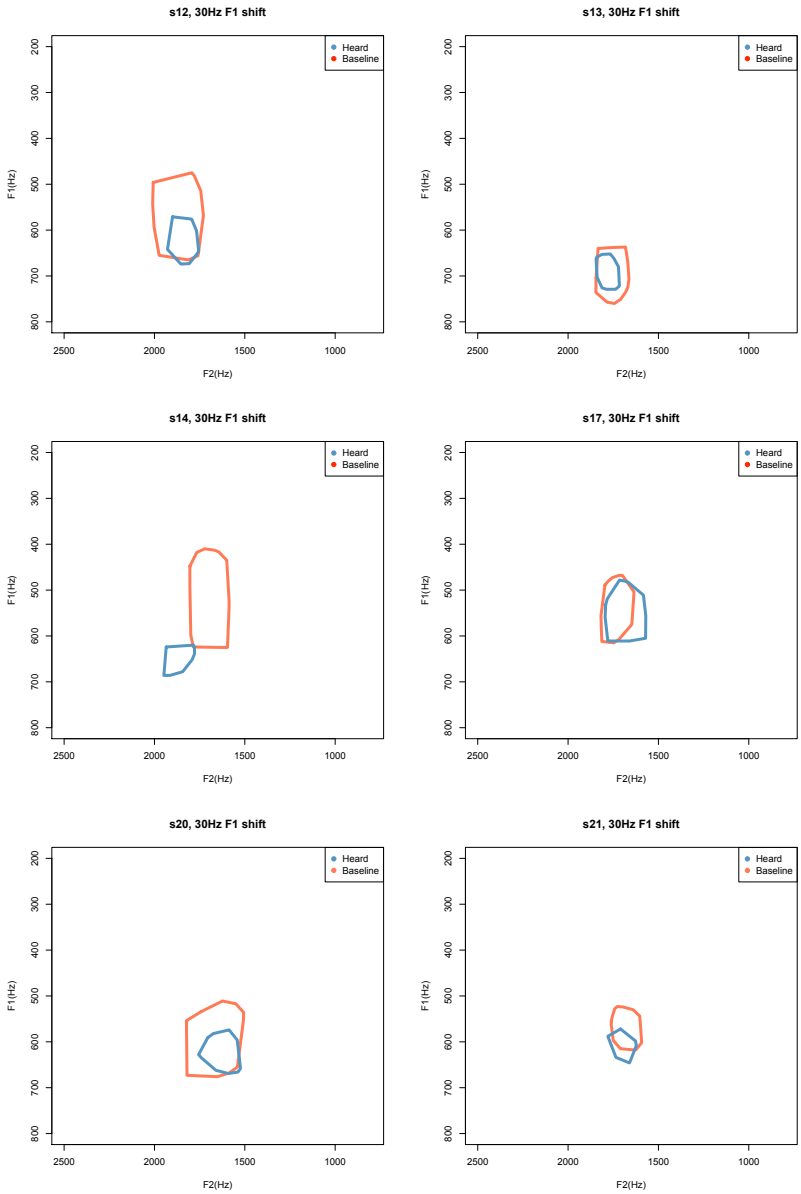


Figure A.1: Response to 30 Hz /ɛ/ F2 feedback shift.

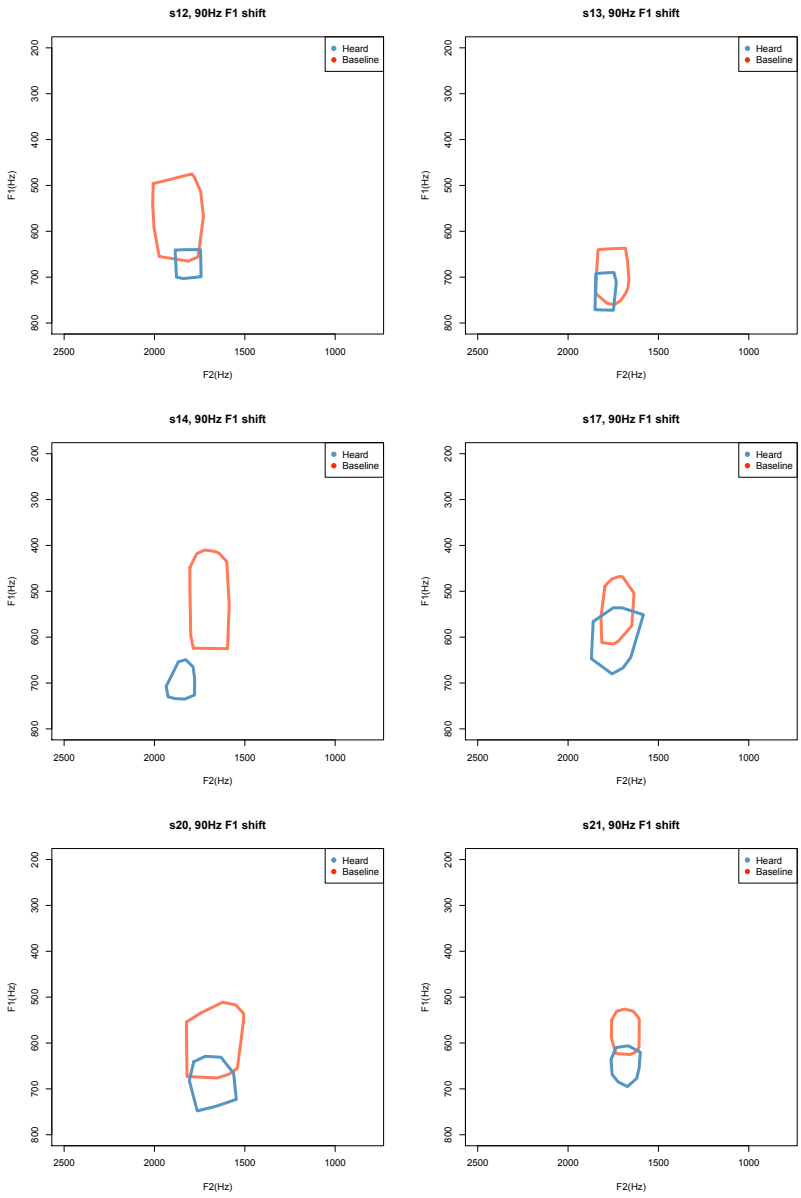


Figure A.2: Response to 90 Hz /ɛ/ F2 feedback shift.

Appendix B

For completeness, the results of linear regression analyses predicting F1 and F2 for individual experiments with citation form words used to predict baseline vowel frequencies are listed below.

F1 ɛ to ɪ	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	0.067	0.045	1.49	0.139
completeness	0.053	0.036	1.46	0.146
Pr(produce intended V)	0.051	0.037	1.37	0.174
previous F1	0.50	0.065	7.73	7.88e-13 ***

Table B.1: Predictors in linear regression predicting F1 in citation vowels.

F1 ɛ to æ	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	0.068	0.012	5.58	8.99e-08 ***
completeness	0.032	0.031	1.06	0.293
Pr(produce intended V)	-0.048	0.019	-2.52	0.0128 *
previous F1	0.24	0.056	4.23	3.70e-05 ***

Table B.2: Predictors in linear regression predicting F1 in citation vowels.

F1 ɛ -250Hz	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	-0.13	0.045	-2.96	0.00362 **
completeness	0.090	0.046	1.97	0.0511 .
Pr(produce intended V)	-0.073	0.052	-1.41	0.162
previous F1	0.83	0.051	16.4	$<$ 2e-16 ***

Table B.3: Predictors in linear regression predicting F1 in citation vowels.

F1 ʌ -250Hz	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	0.15	0.097	1.58	0.117
completeness	-0.027	0.034	-0.80	0.427
Pr(produce intended V)	0.17	0.053	3.24	0.00143 **
previous F1	0.84	0.037	22.6	$<$ 2e-16 ***

Table B.4: Predictors in linear regression predicting F1 in citation vowels.

F1 u +300Hz	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	0.26	0.37	0.69	0.490
completeness	-0.011	0.099	-0.11	0.914
Pr(produce intended V)	-0.038	0.18	-0.21	0.831
previous F1	0.88	0.084	10.5	7.31e-16 ***

Table B.5: Predictors in linear regression predicting F1 in citation vowels.

F2 ɛ to ɪ	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	-0.034	0.023	-1.46	0.146
completeness	0.062	0.019	3.33	0.00105 **
Pr(produce intended V)	-0.0090	0.020	-0.457	0.648
previous F2	0.97 0.017	57.2	$<$ 2e-16 ***	

Table B.6: Predictors in linear regression predicting F1 in citation vowels.

F2 ɛ to æ	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	-0.022	0.012	-1.80	0.0737 .
completeness	-0.30	0.051	-5.90	1.84e-08 ***
Pr(produce intended V)	0.063	0.019	3.22	0.00151 **
previous F2	0.66	0.048	13.8	$<$ 2e-16 ***

Table B.7: Predictors in linear regression predicting F1 in citation vowels.

F2 ɛ -250Hz	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	0.011	0.025	0.45	0.654
completeness	0.073	0.024	3.06	0.00270 **
Pr(produce intended V)	0.078	0.028	2.764	0.00651 **
previous F2	0.62	0.062	9.87	$<$ 2e-16 ***

Table B.8: Predictors in linear regression predicting F1 in citation vowels.

F2 ʌ -250Hz	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	0.13	0.17	0.80	0.427
completeness	0.0048	0.056	0.087	0.931
Pr(produce intended V)	-0.036	0.093	-0.39	0.700
previous F2	0.88	0.037	23.6	$<$2e-16 ***

Table B.9: Predictors in linear regression predicting F1 in citation vowels.

F2 u+300Hz	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	0.43	0.56	0.77	0.442
completeness	-0.11	0.13	-0.84	0.406
Pr(produce intended V)	-0.53	0.28	-1.93	0.0581 .
previous F2	0.075	0.12	0.65	0.517

Table B.10: Predictors in linear regression predicting F1 in citation vowels.

F1 ɛ to ɪ	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	-0.34	0.023	-14.8	< 2e-16 ***
completeness	0.13	0.038	3.36	0.000955 ***
Pr(produce intended V)	0.13	0.020	6.26	2.83e-09 ***
previous F1	0.29	0.043	6.76	1.98e-10 ***

Table B.11: Predictors in linear regression predicting F1 in casual speech.

F1 ɛ to æ	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	-0.295	0.026	-11.3	< 2e-16 ***
completeness	0.011	0.02991	0.38	0.707
Pr(produce intended V)	0.090	0.019	4.86	2.55e-06 ***
previous F1	0.20	0.053	3.85	0.000166 ***

Table B.12: Predictors in linear regression predicting F1 in casual speech.

The complete results from linear regression analyses predicting F1 and F2 with background based on the 30-minute interview are listed below.

F1 ɛ -250Hz	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	-0.38	0.047	-8.18	1.82e-13 ***
completeness	0.0025	0.030	0.081	0.935
Pr(produce intended V)	0.060	0.025	2.40	0.0179 *
previous F1	0.54	0.056	9.68	< 2e-16 ***

Table B.13: Predictors in linear regression predicting F1 in casual speech.

F1 ʌ -250Hz	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	-0.0082	0.020	-0.40	0.690
completeness	-0.0097	0.020	-0.481	0.631
Pr(produce intended V)	-0.43	0.046	-9.43	<2e-16 ***
previous F1	0.64	0.051	12.5	<2e-16 ***

Table B.14: Predictors in linear regression predicting F1 in casual speech.

F1 u+300Hz	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	0.015	0.080	0.18	0.85
completeness	0.19	0.079	2.37	0.0207 *
Pr(produce intended V)	-0.15	0.10	-1.44	0.15
previous F1	0.98	0.061	16.1	<2e-16 ***

Table B.15: Predictors in linear regression predicting F1 in casual speech.

F2 ɛ to ɪ	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	-0.020	0.019	-1.03	0.304
completeness	-0.011	0.033	-0.33	0.742
Pr(produce intended V)	-0.022	0.017	-1.26	0.208
previous F2	0.98	0.018	55.8	<2e-16 ***

Table B.16: Predictors in linear regression predicting F2 in casual speech.

F2 ɛ to æ	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	-0.0031	0.031	-0.10	0.921
completeness	-0.075	0.043	-1.74	0.0838 .
Pr(produce intended V)	0.059	0.023	2.53	0.0123 *
previous F2	0.78	0.047	16.7	<2e-16 ***

Table B.17: Predictors in linear regression predicting F2 in casual speech.

F2 ɛ -250Hz	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	-0.11	0.024	-4.64	8.00e-06 ***
completeness	0.011	0.015	0.70	0.487
Pr(produce intended V)	0.091	0.015	6.07	1.20e-08 ***
previous F2	0.37	0.062	5.91	2.67e-08 ***

Table B.18: Predictors in linear regression predicting F2 in casual speech.

F2 ʌ -250Hz	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	-0.14	0.043	-3.32	0.00109 **
completeness	0.0048	0.038	0.12	0.901
Pr(produce intended V)	0.103	0.088	1.18	0.241
previous F2	0.76	0.059	12.9	$<$ 2e-16 ***

Table B.19: Predictors in linear regression predicting F2 in casual speech.

F2 u +300Hz	**Estimate**	**Std. Error**	**t value**	**Pr**$(>\|t\|)$
density	0.016	0.13	0.12	0.902
completeness	0.19	0.12	1.65	0.103
Pr(produce intended V)	-0.50	0.16	-3.13	0.00254 **
previous F2	0.10	0.11	0.98	0.331

Table B.20: Predictors in linear regression predicting F2 in casual speech.