Visual feedback during speech production

by

Elizabeth Leigh Stelle

BA (Hons), The University of Queensland, 2006

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL
STUDIES
(Linguistics)

The University Of British Columbia
(Vancouver)

December 2016

© Elizabeth Leigh Stelle, 2016
Abstract

The visual speech signal has a well-established influence on speech perception, and there is growing consensus that visual speech also influences speech production. However, relatively little is known about the response to one’s own visual speech; that is, when it is presented as speech feedback. Since visual feedback is generated by the same speaking event that generates auditory and somatosensory feedback, it is temporally compatible with these typical sources of feedback; as such, it is predicted to influence speech production in comparable ways. This dissertation uses a perturbation paradigm to test the effect visual feedback has on production.

Two delayed auditory feedback experiments tested the effect of different types of visual feedback on two fluency measures: utterance duration and number of speech errors. Visual feedback was predicted to enhance fluency. When the presentation of static and dynamic visual feedback was randomized within a block, utterance duration increased with dynamic visual feedback but there was no change in speech errors. Speech errors were reduced, however, when the different types of visual feedback were presented in separate blocks. This reduction was only observed when dynamic visual feedback was paired with normal auditory feedback, and for those participants who were more verbally proficient. These results suggest that consistent exposure to visual feedback may be necessary for speech enhancement, and also that the time-varying properties of visual speech are important in eliciting changes in speech production.

In the bite block experiment, participants produced monosyllabic words in conditions that differed in terms of the presence or absence of visual feedback and a bite block. Acoustic vowel contrast was enhanced and acoustic vowel dispersion was reduced with visual feedback. This effect was strongest at the beginning of the
vowel and tended to be stronger during productions without the bite block. For a small subset of participants the magnitude of motion of the lower face increased in response to visual feedback, once again without the bite block.

The results of this dissertation provide evidence that visual feedback can enhance speech production, and highlight the multimodal nature of speech processing.
Preface

This dissertation is original work by the author, Elizabeth Leigh Stelle.

The experiments in this dissertation were run under approval of The University of British Columbia Behavioural Research Ethics Board, (certificate no. H12-02559) in Canada, and The University of New Mexico Institutional Review Board (certificate no. 12-587) in the USA. The images in Figure 2.1, Figure 4.3, and Figure 4.5 are used with the consent of the participants.

Parts of Chapter 3 have been published as Stelle, E., Smith, C.L., & Vatikiotis-Bateson, E. Delayed auditory feedback with static and dynamic visual feedback. In Proceedings of The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing, 40-45. The lead author (Stelle) designed and implemented the experiment, analyzed the data, and wrote the manuscript. Writing of the manuscript involved input from the additional authors (Vatikiotis-Bateson, Smith).
Table of Contents

Abstract ......................................................... ii
Preface ......................................................... iv
Table of Contents ........................................... v
List of Tables .................................................. viii
List of Figures .................................................. x
Acknowledgements ........................................... xii
Dedication ....................................................... xiv

1 Introduction ................................................. 1
  1.1 Feedback during speech production ......................... 3
  1.1.1 Feedback and feedforward motor control ............... 5
  1.1.2 Altered speech feedback .................................. 6
  1.1.3 Visual speech feedback .................................. 9
  1.2 On the use of novel feedback ............................... 14
  1.3 Outline of the dissertation ................................. 16

2 Methods ....................................................... 19
  2.1 Overview .................................................. 19
  2.2 Equipment ............................................... 19
  2.3 Stimuli ................................................... 20
  2.4 Procedure ............................................... 21
4.3.4 Measures ........................................ 88
4.4 Results ............................................. 95
  4.4.1 Vowel contrast ................................. 97
  4.4.2 Vowel dispersion ............................. 101
  4.4.3 Lower face magnitude of motion .......... 108
4.5 Discussion ......................................... 116

5 Discussion ......................................... 123
  5.1 Overview .......................................... 123
  5.2 Summary of experimental results ............... 123
  5.3 Multimodal speech production .................. 126
    5.3.1 Targets of production ...................... 126
    5.3.2 Multimodal feedback integration .......... 128
    5.3.3 Summary .................................. 134
  5.4 Future work ..................................... 134

Bibliography ........................................ 137

A Stimuli ............................................. 153
List of Tables

Table 3.1  Conditions presented to each participant. .................. 40
Table 3.2  Disfluency transcription conventions from Brugos and Shattuck-
Hufnagel (2012). .......................................................... 42
Table 3.3  Cross-tabulation of error categories from each coder. ...... 45
Table 3.4  Means and standard errors of the means of the experimental
measures for each condition. ................................. 45
Table 3.5  Fixed effects for the model of (log) utterance duration. ...... 47
Table 3.6  Fixed effects for the model of speech errors. .................. 50
Table 3.7  Fixed effects for the model of speech errors with the significant
effect of DAF disruption. ............................................. 51
Table 3.8  Conditions presented to each participant. .................... 58
Table 3.9  Cross-tabulation of error categories from each coder. ...... 59
Table 3.10  Means and standard errors of the means of the experimental
measures for each condition. ................................. 60
Table 3.11  Fixed effects for the model of (log) utterance duration with a
significant interaction of the predictors. .......................... 62
Table 3.12  Fixed effects for the model of speech errors with the significant
auditory feedback effect. ............................................... 65
Table 3.13  Fixed effects for the model of speech errors with the signif-
icant interaction between auditory feedback, visual feedback, and
DAF disruption. ......................................................... 66
Table 3.14  Number of pauses per 100 sentence for each condition in Ex-
periment 1 and Experiment 2. ................ ............... 74
Table 3.15 Means and standard errors of the means for the dependent variables in Experiment 1 and Experiment 2. 75

Table 4.1 The four conditions presented to each participant. 86

Table 4.2 The four condition orders used in the experiment. 88

Table 4.3 Summary of the dependent variables and effects structures used in the statistical models of vowel contrast. 100

Table 4.4 Fixed effects for the model of (cube root) AVS (25%) with the significant visual feedback effect. 100

Table 4.5 Fixed effects for the model of (cube root) AVS (50%) with a near-significant interaction of the predictors. 101

Table 4.6 Summary of the dependent variables and effects structures used in the statistical models of vowel dispersion. 103

Table 4.7 Fixed effects for the model of (cube root) Euclidean distance (25%) with the significant visual feedback and phoneme effects. 104

Table 4.8 Fixed effects for the model of (cube root) Euclidean distance (50%) with the significant interaction between type of oral perturbation and type of visual feedback. 105

Table 4.9 Fixed effects for the model of (cube root) Euclidean distance (75%) with the significant effect of phoneme. 106

Table 4.10 Fixed effects for the model of (log) magnitude of motion. The effect of oral feedback was close to significance. 110

Table A.1 Stimuli used in the DAF experiments (Chapter 3). 157
Table A.2 Stimuli used in the bite block experiment (Chapter 4). 159
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>An example of the visual feedback presented to a participant.</td>
<td>20</td>
</tr>
<tr>
<td>2.2</td>
<td>Simulated data demonstrating homoscedasticity (left) and heteroscedasticity (right).</td>
<td>28</td>
</tr>
<tr>
<td>3.1</td>
<td>An example of a complex disfluent event (error code: e.ps).</td>
<td>43</td>
</tr>
<tr>
<td>3.2</td>
<td>An example of a part-word prolongation.</td>
<td>44</td>
</tr>
<tr>
<td>3.3</td>
<td>Distributions of utterance duration for each condition.</td>
<td>46</td>
</tr>
<tr>
<td>3.4</td>
<td>Counts of each type of error per condition.</td>
<td>48</td>
</tr>
<tr>
<td>3.5</td>
<td>Density plots of speech error counts for each condition.</td>
<td>49</td>
</tr>
<tr>
<td>3.6</td>
<td>Effect of DAF disruption on speech errors.</td>
<td>52</td>
</tr>
<tr>
<td>3.7</td>
<td>Distributions of utterance duration for each condition.</td>
<td>61</td>
</tr>
<tr>
<td>3.8</td>
<td>Counts of each type of error per condition.</td>
<td>63</td>
</tr>
<tr>
<td>3.9</td>
<td>Density plots of speech error counts for each condition.</td>
<td>64</td>
</tr>
<tr>
<td>3.10</td>
<td>Interaction between DAF disruption, auditory feedback, and visual feedback.</td>
<td>67</td>
</tr>
<tr>
<td>4.1</td>
<td>Predicted effects of the experimental manipulations on the variables to be measured.</td>
<td>84</td>
</tr>
<tr>
<td>4.2</td>
<td>The canonical positions in the vowel space for the General American English monophthongs used in the experiment (adapted from Ladefoged (2006)).</td>
<td>85</td>
</tr>
<tr>
<td>4.3</td>
<td>Bite block configuration used in the oral perturbation conditions.</td>
<td>87</td>
</tr>
<tr>
<td>4.4</td>
<td>Mahalanobis distances from a token of the vowel [i] to the distribution of each vowel category.</td>
<td>92</td>
</tr>
</tbody>
</table>
Figure 4.5  A screenshot from the FlowAnalyzer software with a region of interest (ROI; black rectangle) specified for a particular participant. ................................................................. 93
Figure 4.6  Comparison of mouth and lower face magnitudes of motion from two ROIs. ................................................................. 95
Figure 4.7  Vowel space as measured at the 25% point in the vowels. . . 96
Figure 4.8  Vowel space as measured at the 50% point in the vowels. . . 96
Figure 4.9  Vowel space as measured at the 75% point in the vowels. . . 97
Figure 4.10  Means and standard errors for AVS distance for each condition at three points in the vowel. ........................................ 98
Figure 4.11  Means and standard errors for Euclidean distance for each condition at three points in the vowel. ............................... 102
Figure 4.12  Means and standard errors for Euclidean distance for each phoneme in each condition at each measurement point in the vowel. . . 107
Figure 4.13  Means and standard errors for magnitudes of motion of the lower face for each condition. .......................... 108
Figure 4.14  Different patterns of magnitudes of motion of the lower face in response to visual feedback. ................................. 110
Figure 4.15  Non-high vowels: Magnitudes of motion of the lower face for each participant’s production in each condition. ................ 112
Figure 4.16  High vowels: Magnitudes of motion of the lower face for each participant’s production in each condition. ............. 113
Figure 4.17  Correlations between mean MM and mean AVS. ............ 115
Figure 4.18  Results from Section 4.4.1 and Section 4.4.2 re-plotted to show how the relationship between vowel contrast (top) and vowel dispersion (bottom) changes throughout the vowel. ........... 118
Acknowledgements

Without question, the best part of my graduate school experience has been the people I’ve met along the way.

My co-advisors, Eric Vatikiotis-Bateson and Caroline Smith, were in many ways a perfect complement to one another. Eric gave me free rein to pursue my research interests while Caroline kept me focused on actually finishing my dissertation. Many thanks go to Eric for being an unflappable advisor who offered forthright guidance and genuine kindness. I am immensely grateful to Caroline for taking me on as a student; not only was she generous with her time and thoughtful with her feedback, but she also made it possible for me to run my experiments at UNM. I couldn’t have finished this work without her. I was fortunate to have two other committee members for my dissertation: Molly Babel and Bryan Gick. Molly’s attention to detail helped me to think more carefully about my work and Bryan encouraged me to consider the bigger picture. My defense was a more energizing and exciting experience than I had anticipated. In addition to my committee, I’d like to thank my university examiners, Kathleen Curry Hall and Janet Werker, and my external examiner, Jeffery Jones, for their insightful questions and engaging discussion.

I had assistance from many grad student colleagues while running my experiments. Thanks go to Michael Fry for error coding, Michael McAuliffe for sharing his OFA Python script, and Alexis Black, Megan Keough, Michael McAuliffe, Ella Fund-Reznicek, and George Stelle for stimuli recordings. I’m not sure I can offer enough thanks to George Stelle for creating all of the experiment software. Thanks also go to the UNM Ling 101 instructors who helped recruit participants. And the UBC Linguistics administrators, Edna Dharmaratne and Shaine Meghji,
helped me in so many ways, especially after I moved to Albuquerque; I offer them my sincerest thanks.

As luck would have it I ended up being part of the very best cohort, alongside Heather Bliss, Elizabeth Ferch, Raphael Girard, Murray Schellenberg, Anita Szakay, and Carmela Toews. After bonding over the intensity of coursework, we’ve remained close despite the fact that we’re now scattered around the globe, and I look forward to our semi-regular reunions for many years to come. Beyond my cohort, I’ve formed many cherished friendships with other grad students in my department which I hold close to my heart. Thank you to Alexis Black, Amelia Reis Silva, and Sonja Thoma for generously welcoming me into their homes during my visits to Vancouver. Alexis was also a steadfast source of encouragement during the final stretch of dissertation writing; I look forward to returning the favor! Mark Scott and Anita Szakay are the kind of friends you meet as an adult but feel like you’ve known your entire life. They are part of my fondest memories of grad school and continue to be important people in my life. Thank you to Mark for talking shop, reading drafts, and listening to practice talks. And thank you to Anita for admonishing me to just get it done!

I’ve made many wonderful friends in Albuquerque, but special thanks go to Jessica Gross, Haley Groves, and Meg Healy for sharing with me the highs and lows of grad school and parenting. I’m grateful for my family and friends in Australia; my parents and siblings have cheered me on through the long years of grad school, as have my dear friends Cate Ryan and Michael O’Brien.

Bertrand Russell described a good life as “one inspired by love and guided by knowledge.” While grad school has deepened my appreciation for critical thinking and scientific inquiry, it is the love I have for my family–George, Adelaide, and Sullivan–which inspires me daily. My children won’t remember any of this, but I’ll remember their role in my dissertation with deep affection, and be forever in awe that we managed to get through it one piece. My husband... Where do I even begin? George, you have been unwavering in your love and support, and have maintained a resolute confidence in me even when I lost confidence in myself. And you’ve done all this while working on your own dissertation. I can’t adequately express my love and admiration, except to say that, by Russell’s measure, we are living the very best life, and I am thankful for it.
Dedication

For George.
Always.
Chapter 1

Introduction

The flexibility we bring to the task of perceiving and producing speech has received considerable attention. Listeners can integrate information from a variety of sensory modalities. For example, conflicting sensory information can result in percepts from a different category than the inputs (e.g. McGurk and MacDonald, 1976) and ambiguous information from one sensory modality can be influenced by information from another modality (e.g. Bertelson et al., 2003). Speakers make use of several feedback channels during the production process, such as auditory (e.g. Houde and Jordan, 1998) and somatosensory (e.g. Tremblay et al., 2003). There is a high degree of redundancy in the speech production-perception system, a feature which is common to most biological systems and which allows for more robust operation, especially in difficult speaking and listening conditions. Such redundancy allows for variable integration of the sensory modalities in different situations.

This dissertation investigates this flexibility by testing speakers’ ability to incorporate atypical speech feedback signals. Specifically, it looks at real-time visual feedback of one’s own speech production. Visual feedback was chosen as a test case for the effects of atypical feedback since, as perceivers, we have extensive experience with this modality. This familiarity may make it more likely that speakers can make use of this information during production. Just as the addition of a visual speech signal can improve perception accuracy, especially in difficult listening conditions (e.g. Navarra and Soto-Faraco, 2007; Sumby and Pollack, 1954), visual feedback is hypothesized to improve speech production in terms of measures such
as fluency and phonemic contrast. The expectation is that visual feedback will have the strongest effect in difficult speaking conditions, in a parallel manner to the improvement found in perception when a visual signal is added to difficult listening conditions, such as speech in noise. Indeed, there is evidence to suggest that speech production stimulated by an audiovisual signal is more accurate in shadowing tasks than an audio-only signal (Reisberg et al., 1987; Scarbe et al., 2014) and more fluent for aphasic patients (Fridriksson et al., 2015, 2012). Finally, visual feedback has temporal properties that make compatibility with other sources of speech feedback likely, since all sources of feedback are generated from the same act of speaking.

This dissertation presents evidence in support of this hypothesized role for visual feedback during speech production. The experiments reported here use a perturbation paradigm to create difficult speaking conditions in which to compare the presence and absence of visual feedback. Two types of perturbations are used: delayed auditory feedback (auditory perturbation) and bite block speech (somatosensory perturbation). The visual feedback comparison is also made for non-perturbed speech in each experiment. The results of the delayed auditory feedback experiments suggest that it is the time-varying properties of visual feedback, as opposed to the static form properties, that can globally influence the production of whole utterances, but these changes are reduced when the presentation of dynamic and static visual feedback is randomized rather than presented in consistent blocks (Chapter 3). The results of the bite block experiment suggest that changes to production can also be observed in individual segments; vowels exhibit greater acoustic contrast and reduced dispersion when produced with visual feedback. Acoustic contrast is also modestly correlated with lower face movement (Chapter 4).

Before moving forward, a terminological clarification is needed. The term visual feedback is used variably in the literature. It is often used to refer to the visualization of an acoustic or articulatory property of speech, such as formants (Kartushina et al., 2015), pitch (de Bot, 1984), or jaw position (Loucks and De Nil, 2006). Visual feedback can also be used to refer to biofeedback of non-visible articulators, such as the tongue, by way of ultrasound (Adler-Bock et al., 2007) or electromagnetic midsagittal articulography (Katz and Mehta, 2015). In this dissertation, visual feedback refers to the front view of one’s face, as one would see
when looking in a mirror. This is the same view one typically has of an interlocu-
tor, at least when talking face-to-face, and is also consistent with the type of visual
stimuli commonly used in audiovisual perception research.

The rest of this chapter consists of a general overview of feedback, including
visual feedback, during speech production, taking into account its interaction with
feedforward control and considering evidence for the importance of feedback based
on experiments using altered speech feedback.

1.1 Feedback during speech production

Feedback refers to information about the sensory consequences of an action that is
fed back to the motor control system in order to monitor and, if necessary, change
the action. Actions can be changed online or the feedback can be used to train
the system and thus change subsequent actions. Three types of sensory feedback
that are normally available during speech production include auditory, tactile, and
proprioceptive feedback. Auditory feedback is received via both air- and bone-
conducted sound. Tactile feedback includes contact between articulators as well as
changes in air pressure that are registered as the sensation of touch by the vocal
tract walls. Proprioceptive feedback provides information about the movements
and position of speech-relevant body structures. Interestingly, while the oral cavity
and lips have very high tactile sensitivity (Miller, 2002), the oro-facial muscles are
largely devoid of proprioceptors (Cattaneo and Pavesi, 2014). Instead, mechanore-
ceptors in the skin, responding to the stretch of skin due to the muscle activity
beneath it, most likely provide proprioceptive feedback (Ito and Ostry, 2010).

An alternative term for proprioceptive feedback that is commonly used is (so-
mato)sensory feedback. Early bite block studies referred to the feedback from
articulators as “sensory feedback” (e.g. Fowler et al., 1980; McFarland and Baum,
1995; McFarland et al., 1996), with later studies using “somatosensory” (e.g. Lane
et al., 2005) or “oro-sensory” (e.g. Namasivayam et al., 2009) as a cover term
for both tactile and proprioceptive feedback. Perturbation studies which involve
mechanically altering jaw trajectories often refer to this movement information as
“somatosensory feedback” (e.g. Feng et al., 2011; Lametti et al., 2012; Nasir and
Ostry, 2008; Tremblay et al., 2003). In this dissertation, feedback concerning the
movement of and contact between articulators is particularly relevant to the bite block experiment in Chapter 4. In keeping with the terminological practices most common in these types of studies, somatosensory feedback will be used to refer to tactile and proprioceptive feedback.

As stated above, feedback can be used for the control of actions. Control systems are classically categorized as either open-loop control and closed-loop control systems, a distinction based on the use of feedback (e.g. Hood, 1998). Open-loop control is not context-dependent; external events do not affect the way a goal is achieved and feedback is not used to ensure a successful end state. In contrast, closed-loop control is context-dependent. External information is used to achieve the desired end-state, thus the control system makes adjustments to its performance during the task.

A simple example is a clothes dryer, a device that can function using either type of control system. Using the timer setting on the dryer is an example of open-loop control. The duration is pre-set and the drying cycle will run to completion even if the clothes become dry before the cycle is complete. Using the moisture sensor setting on the dryer is an example of closed-loop control. The desired moisture level is set and the moisture sensors monitor the clothes (by monitoring for electrical resistance, for example). The dryer is turned off once the desired moisture level has been reached.

Open-loop control requires minimal computation; however, the more complex closed-loop control is better suited to the adaptive abilities critical to biological systems. An example of this is how the hypothalamus uses negative feedback, or error signals, to regulate body temperature (e.g. by activating sweat glands in response to high body temperature, thus initiating evaporative cooling). There is a problem with closed-loop control, however, and that is that it takes time to generate and process feedback. Feedback from an action is not available immediately; for example, there is a delay of at least 80-100 ms between visual or proprioceptive feedback and its effect on movement, with the delay for reaching movements that have a visual goal being much longer (300-700 ms) (Desmurget and Grafton, 2000). For the motor control of actions to be closed-loop, the various stages involved must be completed within the time frame of a single movement goal, so that the feedback can play its part in ensuring the correct movement is made. These stages involve
the registering of an error, calculation of corrective movements, issuing the motor commands for the corrective movements, and the actual implementation of the correction by the muscles (Perkell, 2012).

This temporal limitation is especially problematic for rapid, highly-skilled activities such as speech. For speech production a range of feedback delays have been reported. Some results suggest that closed-loop motor control may be viable for speech tasks. For example, Tiede et al. (2006) perturbed jaw motion in such a way that the acoustics were consequently changed. They found that evidence for compensation was seen in the acoustic domain before the kinematically perturbed vowel was completed, and beginning as early as 75 ms after the onset of the perturbation. However, longer delays are more commonly reported for speech feedback, with averages closer to 100 ms (e.g. Tourville et al., 2008), 200 ms (e.g. Burnett et al., 1998), or upwards of 400 ms (e.g. Purcell and Munhall, 2006). Thus, the possibility for closed-loop control most likely only arises during very controlled speaking tasks where the speech sounds are sustained, and not during running speech (Perkell, 2012).

Despite all this, there remains a role for feedback during speech production, in the context of feedforward motor control.

1.1.1 Feedback and feedforward motor control

The temporal limitation of closed-loop control systems described in the previous section can be addressed by a system that predicts the sensory consequences of an action in advance of actually receiving the feedback. This type of system is known as a feedforward model, and it involves defining a motor plan before an action is started (Desmurget and Grafton, 2000). Given an initial state, an internal model makes a prediction of the sensory consequences of an action, and this prediction can be used as feedback before the actual—or ‘reafferent’—sensory feedback is processed. This essentially sidesteps the problem of having to wait for the reafferent signal before updating the motor plan, which is particularly useful for fast sequences of actions.

How, then, does peripheral sensory feedback fit with forward models of motor control? Three important roles for feedback that have been identified include:
learning the internal model (i.e. the motor-sensory relationship), updating the internal model, and handling sudden, or unexpected, perturbations (Hickok, 2014).

Despite the inadequacies of feedback for online motor control, feedback may be important during development in establishing the sensory-motor relationship (Borden, 1979). The importance of this becomes clear when one of the sources of feedback is absent during development. For example, congenitally deaf infants have a considerable delay in the onset of babbling compared to hearing infants (Oller and Eilers, 1988) and the speech of older deaf children with some residual hearing tends to receive low intelligibility scores, the low scores due largely to phoneme production problems (Smith, 1975). In the case of typical development, once the internal model of speech control has reached a mature state, speech production can proceed primarily under the control of feedforward commands.

The next two roles for feedback identified above involve correcting either persistent or sudden errors. In the State Feedback Control model of speech processing, which builds on neurological evidence for sensory predictions of speech (Ventura et al., 2009), Hickok et al. (2011) propose an external monitoring stage. This step of the motor control process considers whether the actual sensory consequences of an articulation match the predicted sensory consequences. When error signals are generated, not only is the motor controller provided with corrective feedback, but the internal model is also fine-tuned so that future errors are minimized. While the neurological evidence for these components of forward models of speech motor control is still evolving, there is considerable behavioral evidence for speakers adjusting their productions in response to errors. Many studies have used a perturbation paradigm to investigate these adjustments by intentionally causing errors. An overview of this research is presented in the next section.

1.1.2 Altered speech feedback

Perturbing a system is one way to investigate how the various parameters contribute to the overall performance. While this rationale has been the starting point for much research, it is not a universally held proposition. As (Borden, 1979, p. 312) argues, “Compensation under abnormal circumstances does not mean peripheral feedback is necessary under normal circumstances.” From this perspective,
altering a component of the system alters the system entirely, thus the response to a perturbation may not provide much insight into the original system. Even with these reservations, it is nevertheless clear that perturbation studies have been a rich source of information on the role of feedback in speech production, with the potential to yield important insight into the coordination of multiple speech signals (e.g. Zimmermann et al., 1988). Numerous studies have explored the effects of altered speech feedback, and a review of acoustic (both spatial and temporal) and somatosensory perturbations is presented below.

In a classic study of this phenomenon in the acoustic domain, Houde and Jordan (1998) played speakers their auditory feedback through headphones while making incremental changes to the formant frequencies. For example, over the course of the training phase of the experiment, participants gradually heard their productions of *pep* sound more like *pip*, which was achieved by lowering F1. Participants compensated for this feedback shift by making their productions sound more like *pap*, which is produced with a higher F1. In addition to sustained vowels, this effect has been found with longer utterances. Cai et al. (2011) shifted the F2 minimum of [u] in the word *owe* as produced in the phrase *I owe you a yo-yo*. Speakers compensated for the shift by changing their production in a similar fashion to those in the Houde and Jordan study, but to an even greater degree, suggesting that auditory feedback may play a greater role in time-varying gestures compared to sustained vowels.

The nature of this compensatory response to perturbations has been investigated in more detail. As in other studies, Katseff et al. (2012) found that compensation for formant perturbations was incomplete. Interestingly they also observed that small formant perturbations resulted in greater compensation than large formant perturbations. The authors (and others, e.g. Perkell, 2012)) argue that incomplete compensation is due to conflict between auditory and somatosensory feedback; the greater the mismatch between the two feedback signals, the more likely a speaker is to rely on the unperturbed feedback signal, thus minimizing compensation.

The studies just described perturbed spatial properties of the acoustic signal; however, it is also possible to introduce temporal perturbations. Cai et al. (2011) shifted the F2 minimum during the [u] of word *you* as produced in the phrase *I owe you a yo-yo*, either shifting the minimum backward in time (by an average of 45.4
ms) or forward in time (by an average of 24.6 ms). The forward shift, which was described as the “deceleration” condition, resulted in longer durations for several intervals throughout the utterance; the effects of the local perturbation seemed to extend to the F2 landmarks for the whole utterance. In contrast to the compensatory response to spatial perturbations in the acoustic domains, the compensation to the temporal perturbation was in the same direction as the temporal perturbation.

Larger scale temporal shifts can also be used, such as the delayed auditory feedback (DAF) paradigm. In this set up, auditory feedback is typically delayed by 180-200 ms, the effect of which is to impede fluent speech output (Yates, 1963). As with the more focused temporal shifts described in the Cai et al. (2011) study, the response to the perturbation is in the same direction as the perturbation; that is, auditory feedback is shifted forward in time, or ‘decelerated’, and speakers slow their speech in response. This is in contrast to spatial perturbations, which elicit compensations in the opposite direction to the perturbation, as in Houde and Jordan (1998). There is some speculation that the effects of DAF may be exacerbated in speakers who rely more strongly on feedback during production, whether this is due to instability in their feedforward control system (Chon et al., 2013), or a lack of experience with competing auditory inputs (Fabbro and Darô, 1995). This type of perturbation will be discussed in more detail in Chapter 3.

Altered feedback has also been explored in the somatosensory domain. In an early demonstration that perturbations to articulator trajectories elicit a compensatory response, Kelso et al. (1984) perturbed the jaw motion of a speaker by mechanically applying a constant load during the closing gestures of /bæb/ and /bæz/. One of the differences between the final consonants is the relevance of the upper lip; the upper lip is involved in the articulation of /b/ but not /z/. In response the upper lip was lower for the /b/ closure than it was on trials when the jaw was not perturbed. This upper lip lowering was not found for the /z/ closure, which was interpreted as evidence that the jaw, lower lip, and upper lip were acting as a coordinative structure; that is, for the phoneme /b/, there was a “temporary marshaling of many degrees of freedom into a task-specific, functional unit” (Kelso et al., 1984, p. 828).

Tremblay et al. (2003) mechanically altered speakers’ jaw trajectories by applying force in the direction of jaw protrusion. Speakers adapted to this perturba-
tion, shifting their jaw trajectory back to the pre-perturbation path. An after-effect was also produced which was in the opposite direction to the perturbation; that is, the jaw trajectory became retracted. Critically, these perturbations did not measurably affect the acoustic properties of speech, and the compensations were only found when the participants produced speech movements and not when they produced non-speech movements.

Altering the feedback that a speaker receives affects current and subsequent productions, providing evidence for a role for feedback in the control of speech production. This role may be limited in normal speaking conditions, when feedforward control is primarily relied upon, but is nonetheless important in the context of perturbed speech.

1.1.3 Visual speech feedback

The visual speech signal contains a wealth of information. For example, labial cues can be used for vowel identification, and jaw motion can be mapped onto syllables. But visually salient speech information extends beyond the lower face. Motion computed from muscle activity at the lips is correlated with activity at the outer regions of the face (Vatikiotis-Bateson et al., 1996). In an examination of the relationship between vocal tract configurations, facial motion, and speech acoustics, Yehia et al. (1998) demonstrated that orofacial motion is highly correlated with movement of the lips, jaw, and tongue, and also that fairly precise estimates of acoustic parameters (RMS amplitude and line spectrum pairs\(^1\)) can be made from orofacial motion. These correlations between acoustics and facial motion can be further increased when the utterances under examination are produced in the context of more extensive prose (Vatikiotis-Bateson and Yehia, 2002). Abstracting away from facial deformation, rigid body motion of the head has also been found to correlate with speech acoustics; as much as 88% of the variance of speech fundamental frequency can be predicted from head motion (Yehia et al., 2002).

Given these visual cues to speech, it is not surprising that visual information plays an important role in the perception of speech. In the first demonstration of this, Sumby and Pollack (1954) showed improved accuracy for word identifica-

\(^1\)Line spectrum pairs, which are derived from LPC coefficients, map well to vocal tract shapes (Yehia et al., 1998).
tion for audiovisually presented speech in noise compared to audio-only speech in noise. Audiovisual perception research subsequently expanded in many directions. Complementing the correlation between head motion and F0 described above, the perception of speech output from a talking head was found to improve when head motion was added (Munhall et al., 2004a). Audiovisual enhancement for perception was found even with spatially filtered visual signals (Munhall et al., 2004b). The combination of incongruent audio and visual signals was famously found to result in a type of fused percept that is categorized as something different to the auditory and visual signals (McGurk and MacDonald, 1976). The McGurk effect has proved to be very robust; for example, it can be induced with temporally and spatially dislocated audiovisual signals (e.g. Jones and Munhall, 1997; Munhall et al., 1996) and with a gender mismatch between the auditory and visual signals (e.g. Green et al., 1991).

While there is still much that remains unclear in the field of audiovisual speech processing (see Vatikiotis-Bateson and Munhall (2015)), this brief tour was presented in order to establish that: 1) speech-relevant information is distributed over the whole face, and, 2) as perceivers, we have extensive experience using this information (although which aspects of the visual signal perceivers are using is an open question). Given this, visual speech information is an ideal candidate for the role of ‘atypical speech feedback’; speakers are already familiar with the relation that the visual signal has to speech output, thus making it more likely that it could be used in a novel context.

Visual speech feedback is not something speakers typically have access to. That being said, the possibility of experiencing this kind of feedback is certainly becoming more common due to video chat software, which usually shows each speaker a small video of their own face next to a large video of their interlocutor.2 Visual speech feedback has been the topic of a small, but growing, body of research, which has looked at both non-clinical and disordered speech.

Only a handful of studies have investigated the effects visual feedback has on the speech produced by non-clinical populations. They have all used a similar ex-

---

2The VoIP and videoconferencing software, Skype, is estimated to have generated 2 trillion minutes of video calls in the last 10 years (https://blogs.skype.com/2016/01/12/ten-years-of-skype-video-yesterday-today-and-something-new/).
experimental paradigm involving difficult speaking conditions, created by delaying the auditory feedback. The first of these had an exploratory goal of looking for any effects there might be when visual information was presented during speech production. Tye-Murray (1986) had subjects view their mouth in a mirror while speaking with delayed auditory feedback (DAF), but found no effect of visual feedback on utterance duration (the sole dependent variable that was measured). Using a very similar design, but this time measuring speech errors in addition to utterance duration, Jones and Striemer (2007) found that a subset of participants produced fewer speech errors when they had visual feedback. Most recently, Chesters et al. (2015) looked at both immediate and delayed visual feedback in the context of immediate and delayed auditory feedback. Measuring a range of phonetic properties (such as utterance duration and rhythm), the authors found more disruptive effects of delayed visual feedback than immediate visual feedback (i.e. durations and speech errors were increased). A similar paradigm is used for the experiments in Chapter 3, and these studies are discussed in more detail in that chapter.

In their study looking at speech perception of self and others, Sams et al. (2005) compared a number of conditions using both congruent and incongruent audiovisual stimuli. While this study was not looking at speech production per se, one of the conditions required participants to mouth a syllable while watching themselves in a mirror. This visual signal was paired with an audio signal from another speaker to create an audiovisual stimulus. This contrasted with a condition in which both the audio and visual signals from another speaker. There was no significant difference in syllable identification accuracy between these two conditions; in both cases, participants’ identification of the auditorily presented syllable was adversely affected for incongruent audiovisual stimuli and enhanced for congruent audiovisual stimuli. The authors suggest that the McGurk effect induced in these conditions involves a similar perceptual mechanism, despite the obvious difference that participants were articulating in one condition but not the other.

Building on research into choral speech effects for people who stutter, visual feedback has been proposed as another means for minimizing speech disruptions. Choral speech, or speaking in time with another person, reduces stuttering behaviors, even when a visual-only model (i.e. someone mouthing an utterance) is provided for stutterers to speak in time with (Kalinowski et al., 2000). Snyder
et al. (2009) extended this by providing stutterers with both synchronous and asynchronous (i.e. delayed) visual feedback of their own productions. Both of these experimental conditions resulted in reduced stuttering frequency compared to a control condition with no visual feedback. While Snyder et al. (2009) reported no difference between immediate and delayed visual feedback, Hudock et al. (2011), using a similar experimental design, found a greater reduction in stuttering when visual feedback was delayed.

1.1.3.1 Visually stimulated speech production

Related to the notion of self-generated visual feedback is research looking at speech produced in response to a visual or audiovisual signal from another speaker. One type of evidence that visual speech information can affect productions comes from congenitally blind speakers. Ménard and colleagues (2009; 2013) have documented the differences between blind and sighted French speakers in the production of vowels. Differences were found between the two groups in both the acoustic and articulatory domains. Blind speakers produce smaller acoustic contrast distances between vowel categories (2009; 2013), and greater dispersion within vowel categories (2013). In terms of articulation, blind speakers produced a smaller range of upper lip protrusion (i.e. a visible articulation) and a greater range of tongue backing and tongue curvature (i.e. a non-visible articulation) (2013). These results suggest that visual speech information plays a role in the development of speech targets.

In addition to this evidence from visual deprivation, there is also more direct evidence from the provision of visual speech information. Shadowing tasks require participants to immediately repeat speech. By contrasting audio-only and audiovisual stimuli, these tasks are usually used to explore the role of visual information in speech perception. However, since such studies use speech production output as a dependent variable, they also suggest that visual speech information has a role to play in speech production, a connection also noted by Venezia et al. (2016). Reisberg et al. (1987) found an improvement in terms of percentage of words correctly shadowed, when participants shadowed an audiovisual model compared to an audio-only model. This improvement occurred when shadowing an L2 and a
complex prose passage in the L1. Scarbel et al. (2014) also found an improvement with audiovisual shadowing when using a close-shadowing task, which requires participants to repeat speech as quickly as possible. The experiment compared close-shadowing in noise and clear speech conditions and found that reaction times were faster and responses were more accurate in the audiovisual noise conditions. This was especially so for /apa/ stimuli compared to /ata/ and /aka/.

Venezia et al. (2016) investigated sensorimotor integration of visual speech, looking at the neural regions involved when speech production is stimulated by audio-only, visual-only, and audiovisual inputs. Participants were presented four-syllable strings of CV syllables from the three different types of input, and then they engaged in covert rehearsal of the strings. Covert speech production stimulated by visual-only or audiovisual stimuli was found to involve additional sensorimotor brain regions (the posterior superior temporal region, a multisensory processing area) compared to production stimulated by audio-only. Additionally, regions used during production stimulated by audio-only were activated to a slightly greater extent when the production triggers were visual-only or audiovisual. These results were interpreted as evidence for a dedicated visuomotor speech pathway. This is supported by Fridriksson and colleagues’ (2015; 2012) research showing that patients with damage to their auditory-motor network nevertheless experience enhancement to their speech output when shadowing audiovisual speech.

1.1.3.2 Summary

This foundational research demonstrates that visual speech information, in the form of either self-generated visual feedback or a visual signal from another speaker, can elicit changes in speech production. These changes range from a reduction in speech errors to faster reaction times, and might generally be thought of as enhancements to speech output. In the case of visual feedback, however, the reported effects are quite variable, and as is often the case in the early stages of a research program, more questions are raised than answered. For example, the effects have primarily been tested in the context of disfluent speech, whether artificially or developmentally induced, so it is unknown whether these effects would still be evident in other contexts, or with different stimuli (e.g. words rather than sentences).
There is also the question of which properties of the visual feedback are responsible for the effects; is visual form or visual timing more important? The experiments in this dissertation address these issues, bringing more data to bear on our understanding of multimodal speech processing.

1.2 On the use of novel feedback

One possible objection to this line of research is that it is not ecologically normal; we see other talking faces, but not typically our own. As such, there may be concerns that any changes to speech output that occur in the context of visual feedback tell us little about the more ecologically natural context of speaking without visual feedback. A similar concern was raised in Section 1.1.2 with regard to altered feedback. Two arguments for using visual feedback are presented here.

The use of visual feedback—a source of feedback that is not typically available—is motivated by the fact that speakers have extensive experience seeing their reflection. In addition, there is a precedent for using novel modality pairings in speech perception research, and this research demonstrates that information from the atypical modality can influence perception.

While it is certainly true that visual feedback of one’s own speech is not normal in the majority of communicative contexts, the modern world affords many opportunities for encountering visual speech feedback. Mirrors are ubiquitous, “so cheap and so common that [we] use them to wallpaper our bathrooms and dance floors” (Kleeman, 2016), with many a conversation taking place in their presence. As noted previously, video chat software can provide a view of both the interlocutor and oneself during a conversation. Thus, opportunities for visual speech feedback are perhaps more abundant than one might first think. But despite examples such as these, most people do not have extensive experience speaking at length while attending to visual feedback. Part of the motivation for presenting the whole face, and not just the lower face, as visual feedback in the dissertation experiments, was to match the common experience of looking in a mirror.

Speech perception research sets a precedent for the novel pairing of modalities. This “novelty” can be in terms of the congruity of the signals or the naturalness of the particular modality combination. For example, as described in Section 1.1.3,
the McGurk effect is observed in response to incongruent audio and visual signals, and some research has extended this by introducing additional spatial, temporal, and gender incongruence. In terms of atypical modality pairings, Fowler and Dekle (1991) tested whether haptic speech information, provided by placing a hand on the speaker’s face, could influence auditory speech perception. They found that “feeling” a /ba/ increased the percentage of /ba/ auditory judgements. Similarly, Gick and Derrick (2009) paired auditorily presented voiced and voiceless stops with an aero-tactile signal: mechanically generated air puffs. Perceivers were better at identifying the voiceless consonants when they were accompanied by a puff of air and worse at identifying the voiced consonants when they were paired with the air puff.

The latter two studies are presented within the framework of ecological perception, which emphasizes the role of the environment in structuring information about perceived events (see Fowler (1986) for a detailed outline of direct-realist theories of perception). According to this theory, properties of environmental events (e.g. speaking) can be jointly signaled by different types of environmentally structured energy (e.g. sound, light); in other words, “visible talking and audible talking... are the same event of talking” (Fowler and Dekle 1991, p. 817). This idea of joint specification in the environment is an appealing one, and can easily be extended to rationalize an exploration of novel feedback: visual feedback is structured by the same speaking event that structures auditory and somatosensory feedback, thus all sources of feedback jointly specify the speaking event.

While direct-realism could be a useful starting point for framing predictions about novel speech feedback effects, it introduces theoretical issues that are not addressed in this dissertation; namely, the debate over the objects of perception (e.g. Diehl and Kluender (1989); Fowler (1986); Liberman and Mattingly (1985); and more recently in the context of mirror neurons: e.g. Hickok (2009); Rizzolatti et al. (2001)). A more general interpretation of these types of multimodal effects
is offered by Vatikiotis-Bateson and Munhall (2012). They suggest that the source of information is not critical, so long as the signals have an appropriate temporal pairing and do not contradict ecologically valid multimodal pairings. Framed in this way, visual feedback is predicted to enhance speech output by virtue of the fact that its time-varying properties are compatible with those of the auditory and somatosensory feedback channels, since these different sources of feedback are all generated by the same event of speaking.

Seeing oneself while speaking may not be a normal experience, but the components of that experience are quite normal. Speakers have extensive experience with visual speech from their simultaneous role as perceivers, and quite often have access to reflections of their own face. Even the use of visual feedback to guide actions is normal, albeit for limb motor control (e.g. Sober and Sabes, 2003). While there is the caveat that altering feedback or adding a new signal may limit our insights into the unperturbed system, these research paradigms have nevertheless enriched our understanding of speech processing. Visual feedback has temporal properties that could facilitate alignment with other feedback channels; as such it can be used to investigate speakers’ responses to novel multimodal feedback contexts.

1.3 Outline of the dissertation

This dissertation presents evidence that visual feedback can be used by speakers to enhance their speech output. This evidence comes from two types of perturbation experiments: delayed auditory feedback (DAF) and a bite block perturbation. The results of these experiments are consistent with a small but growing body of research showing that visual information is relevant not only to perception, but also production. Speech produced in the context of another talking face results in greater accuracy during shadowing tasks compared to shadowing an audio-only signal (Reisberg et al., 1987; Scarbel et al., 2014) and greater fluency for aphasic patients (Fridriksson et al., 2015, 2012) and people who stutter (Kalinowski et al., 2000). Visually stimulated speech production has also been shown to involve a dedicated visuo-motor neural pathway, which may underpin the production changes that occur in response to visual speech (Venezia et al., 2016). In the context of
self-generated visual speech feedback, there are indications that speech production is more fluent for stutterers (Snyder et al., 2009) and non-stutterers (Jones and Striemer, 2007) compared to speech that is produced without visual feedback. The experimental results reported in this dissertation provide further evidence for enhancement to speech produced with visual feedback, and also expands the role of visual speech in production by showing that visual feedback is not only relevant for disfluent speech, but also for the fluent production of speech targets. These findings highlight the need for models of speech production to be able to account for a more expanded notion of multimodality; speakers can make adaptive use of information from a range of modalities, including modalities that would not normally be considered ecologically natural, such as visual speech feedback.

A description of the methodological procedures that are common to the experiments in the dissertation is presented in Chapter 2. This chapter also includes an overview of (generalized) linear mixed effects models and details of the model specifications used in the analyses.

The two experiments in Chapter 3 use a DAF paradigm similar to the three previous studies (Chesters et al., 2015; Jones and Striemer, 2007; Tye-Murray, 1986) which have investigated visual feedback during speech production. Experiment 1 takes as its starting point the observation that speakers often use strategies, such as focusing on their articulations, in order to overcome the challenges of producing speech with DAF. The experiment was designed to minimize the opportunities for developing a strategy when producing speech under visual feedback conditions, by varying the order of stimulus presentation. The results suggest that visual feedback does not have a facilitative effect on production in this case, however.

Experiment 2 reinstates the consistent stimulus ordering used in previous research (Chesters et al., 2015; Jones and Striemer, 2007; Tye-Murray, 1986) and expands on the paradigm by contrasting different types of visual feedback: no visual feedback, static visual feedback, and dynamic visual feedback. The time-varying component of the visual feedback, compared to just the static form of the face, was predicted to stimulate greater production changes. The results are somewhat conflicting. While dynamic visual feedback did elicit the greatest changes in production, the predicted facilitatory effect was observed when the dynamic visual feedback was paired with normal, rather than delayed, auditory feedback. When it
was paired with DAF, speech output was slowed to a greater extent than with no visual feedback.

Chapter 4 reports on an experiment in which visual feedback is presented in a previously untested task. In this experiment, a bite block perturbation was introduced, and production measures were made at the level of individual segments rather than whole utterances, as in the DAF experiments. The visual feedback was predicted to enhance the production of vowel contrast and minimize the dispersion of the vowels. Overall, the acoustic results supported both of these predictions, but the effect of visual feedback was mostly evident at the beginning of vowels. Optical flow analysis of the lower face motion revealed considerable interspeaker variation, but a subset of speakers tended to produce greater magnitudes of motion for non-high vowels during unperturbed productions with visual feedback.

Chapter 5 places the experimental results in the context of more general, theoretical issues and presents future directions for this research.
Chapter 2

Methods

2.1 Overview

The experiments reported in this dissertation use a perturbation paradigm to assess the effect visual feedback has on speech production. Chapter 3 reports on delayed auditory feedback experiments and Chapter 4 reports on a bite block experiment. What follows is the methodological procedures these experiments have in common. This chapter describes the equipment used during the experiment and for the recording of stimuli, a basic outline of the experimental procedure, and an overview of the statistical analysis technique.

2.2 Equipment

Stimuli were presented to participants either through Sony MDR-V6 circumaural closed back headphones for the delayed auditory feedback (DAF) experiments or over Altec Lansing ADA215 computer speakers for the bite block experiment. Participants repeated each stimulus, speaking into an Audio-Technica ATR2100-USB Cardioid Dynamic USB microphone, positioned approximately 1-2 inches from the mouth. Participants were video recorded at 30 frames per second (fps) with a Logitech HD Pro Webcam C920 (1080p widescreen), placed on top of a 21.5 inch high-definition LED monitor, approximately 20 inches in front of them. A
custom-written program\(^1\) played the stimuli, audiovisually recorded the speech, fed the audio back to the headphones (with or without a delay) (for the delayed auditory feedback experiments only), and displayed different types of visual feedback on the monitor. In the relevant conditions, participants were able to see their whole face displayed on the monitor and were instructed to maintain this position throughout the experiment. Figure 2.1 shows an example of the visual feedback that participants saw. The audio and video were recorded directly to a laptop computer with sufficient processing power and memory to simultaneously perform the necessary signal processing and also run the experiment.

### 2.3 Stimuli

The lists of sentences and words used in the experiments are presented in Appendix A.

The sentences used as stimuli for the DAF experiments were produced by a female speaker of General American English (Midlands) in a sound attenuating

\(^1\)The program used to run the experiments can be accessed at https://github.com/stelleg/daf.
booth with a Samson CO3U multi-pattern condenser microphone and recorded directly to a computer (16 bit, 44.1 kHz).

The words used as stimuli for the bite block experiment were produced by a male speaker of General American English (Pacific Northwest) in a quiet room with an Audio-Technica ATR2100-USB Cardioid Dynamic USB microphone and recorded directly to a computer (16 bit, 44.1 kHz).

2.4 Procedure

All of the experiments reported in this dissertation used a within-participants design with two independent variables. The primary variable of interest was the type of visual feedback participants received during speech production; visual feedback was either absent, static, or dynamic. The experiments used different subsets of these visual feedback options. When visual feedback was absent participants looked at a fixation point on the monitor. Static visual feedback was a still image of the participant’s face taken before the experiment, and dynamic visual feedback was a real-time video of the participant producing the utterance.

The second independent variable manipulated a non-visual source of feedback; either auditory feedback (Chapter 3: delayed auditory feedback) or somatosensory feedback (Chapter 4: bite block). The specifics of these variables are discussed in the relevant chapters.

Prior to the bite block experiment, participants were shown the words that would be presented auditorily in the experiment, along with sample sentences containing each word (Appendix A). Since these words would be repeated throughout the experiment, a preview was provided to ensure that participants knew what the target form was. Each experiment started with a practice block in which participants repeated an auditorily presented set of sentences (Chapter 3) or words (Chapter 4) that were not part of the main experiment. Participants produced speech in the practice blocks while looking at a fixation point on the monitor, and without any auditory or somatosensory perturbations. The auditory presentation of the stimuli throughout the experiments was self-paced, and participants were given an opportunity to take a break between each condition. Participants were instructed to repeat each sentence or word as accurately as possible. For the DAF experiments
they were also instructed to avoid re-starting a sentence if they made an error. For all of the experiments, participants were instructed to look at the monitor while repeating each sentence or word.

After data collection was complete, the video recordings from the DAF experiments revealed that many of the participants were not looking at the monitor for a substantial number of the stimuli. In order to track the effect this might have on the results, each utterance was coded for whether or not the participant was looking at the monitor while they produced the utterance. This coding was based solely on qualitative judgments that were made according to the following criteria.

1. Before the experiment began participants were required to look at their face on the monitor so that a still image could be captured for use in the experiment. The region that participants focused on when they took the still image of their face was classified by the coder as the “looking region.”

2. For any given item, a participant was counted as looking at the monitor if their focus stayed in the looking region for at least half of the time they were speaking.

This information—which will be referred to as the Looking variable—was included in the statistical analyses for the DAF experiments to control for any effect it might have on the dependent variables.

A review of the videos from the bite block experiment showed that participants consistently looked at the monitor throughout the experiment, so the looking variable was not included in the analyses for that experiment.

2.5 Measures

Different measures were used across the experiments. The dependent variables for the DAF experiments were utterance duration and number of speech errors per utterance (Section 3.3.2.4). For the bite block experiment the dependent variables included acoustic measures (Euclidean distance and average vowel space (calculated from Mahalanobis distances)) and articulatory measures (mean magnitude of motion of the lower face) (Section 4.3.4). The specifics of each measure are discussed in the relevant chapters.
2.6 Statistical analyses

For each experiment, the statistical significance of the results was determined by (generalized) linear mixed effects analyses, a class of regression analysis that incorporates both fixed and random effects. This type of statistical model is particularly useful for repeated measures experimental designs and has the advantage of being robust in the face of missing data points. For each dependent variable reported in this dissertation, a default model was first constructed, and on the basis of subsequent model criticism, any necessary changes were made. Model criticism (or, model validation) assesses whether the model assumptions have been met, and the procedure for checking this is described below in Section 2.6.4. The default model specifications are described here and any experiment-specific changes to this default are discussed in the relevant chapters.

The analyses were implemented in R (R Core Team, 2014) using either the lme4 package (Bates et al., 2014) or the glmmADMB package (Skaug et al., 2014). Models were constructed to investigate the relationship between each dependent variable and the two independent variables: type of visual feedback and type of non-visual feedback. The levels of the independent variables were treatment coded and entered into the model as fixed effects with an interaction term, and the looking control variable was also added. The maximal random effects structure justified by the design was used. Random effects included intercepts for participants and items, as well as by-participant and by-item random slopes for the effect of the interaction between type of visual feedback and type of non-visual feedback (more details on the random effects structure are given below in Section 2.6.2).

While there are several ways to compute p-values for the fixed effects, the simulations reported by Barr et al. (2013) suggest that likelihood ratio tests are one of the best methods. This model comparison approach was used for the analyses presented here. The log likelihood of the full model with the effect in question was compared to the log likelihood of a model without this effect, using the anova() function in R.

Winter (2013) presents an accessible conceptual overview of mixed effects models. For the analysis in bite block experiment (Chapter 4), the phoneme variable was included as a random effect instead of the item variable, since phonemes and items were conflated in the stimulus set (i.e. each phoneme category was represented by a single word).
### 2.6.1 Generalized linear mixed effects models

While a linear model predicts the dependent variable as a linear function of the independent variables, a generalized linear model takes this linear model and relates it to the dependent variable by way of a link function. A link function is “a continuous function that defines the response of variables to predictors in a generalized linear model, such as logit and probit links. Applying the link function makes the expected value of the response linear and the expected variances homogeneous.” (Bolker et al., 2009, p. 128) This is necessary when the dependent variable has certain distributional properties; for example, when it is a discrete probability distribution such as the binomial distribution. A generalized linear mixed effects model additionally allows for the specification of both fixed and random effects.

The counts of speech errors in the DAF experiments (Chapter 3) conform to a Poisson distribution, and as such are modeled using a generalized linear mixed effects model with a logarithmic link function. The log link provides the relationship between the linear predictors and the mean of the Poisson distribution. Since the data contained more zero counts than expected based on the Poisson distribution (Zuur et al., 2009), zero-inflation was specified in the model to avoid biased parameter estimates and standard errors. The fixed and random effects were the same as those outlined above in Section 2.6.

### 2.6.2 Random effects

Random effects are “factors with levels randomly sampled from a much larger population,” in comparison with fixed effects which are “factors with repeatable levels” (Baayen, 2008, p. 241). In linguistics experiments, the most common factors included as random effects are participants and stimuli, and that is the case in the experiments presented here. The random effects structure of the model can specify both an intercept and a slope. The intercept allows the model to make adjustments to the dependent variable on the basis of each level of the random factor. For example, participants will vary in their baseline speech rate, and the model can capture this by adjusting the intercept for each participant. The slope allows the independent variable to have different effects on the levels of the random factor. For example, participants will vary in the effect that delayed auditory feedback has
on their speech rate; some participants will only reduce their speech rate a little but others will exhibit substantial decreases in speech rate.\footnote{Note that the intercept and slope parameters that get included in the model are not the individual adjustments for each level of the random factors, but the overall variance of the random factors. The individual adjustments, known as Best Linear Unbiased Predictors, can be calculated from the random effects parameter estimates.}

It is possible to use a data-driven approach to the specification of the random effects structure. In this case, likelihood ratio tests are used to determine whether a particular random effects parameter improves the model fit; if it does not, then it is removed (Baayen et al., 2008). An alternative to this is the design-driven approach of Barr et al. (2013). In this case, the maximal random effects structure justified by the experimental design is specified, in keeping with the traditional standard for mixed-model ANOVAs to categorize both participants and items as random effects. For the experiments in this dissertation that includes both random intercepts and random slopes, as specified in the default model described above. Using simulated data, Barr et al. (2013) showed that such fully specified models do not suffer from a significant loss of power compared to models in which the random effects have been simplified on the basis of model comparison, and they also have the advantage of minimizing Type I error rates. Given this, and the fact that linear mixed effects models are being used in this dissertation for confirmatory hypothesis testing rather than data exploration, the maximal approach to specifying random effects was used.

This approach is not without problems, however. Sometimes such a model will not converge if there is insufficient data to support the number of parameters introduced by such a complex model. The strategy used in this dissertation to deal with this situation is to simplify the random effects structure in a stepwise manner until the model converges. For example, take as a starting point the maximal model in Equation 2.1, which has been presented using the \texttt{lmer} package syntax. The first line includes the function call, \texttt{lmer()}, and specifies that the dependent variable (DV) is modeled as depending on an interaction between two independent variables (IV\textsubscript{1} and IV\textsubscript{2}). This is the fixed effects portion of the model. The next two lines specify the random effects structure. To the right of the vertical line are the grouping factors (participant and item) and to the left of the vertical line are
the specifications for the grouping factors. The 1 specifies that there is a random intercept for participant and item. The independent variables (IV1 * IV2) specify the way in which the random slope can vary.

\[
\text{lmer(DV} \sim \text{IV1} \ast \text{IV2} + \\
(1 + \text{IV1} \ast \text{IV2} \mid \text{participant}) + \\
(1 + \text{IV1} \ast \text{IV2} \mid \text{item}), \text{data})
\]

(2.1)

If this model fails to converge, the first simplification would be to remove the interaction term for the random slope. Since there is usually less variance in experimentally-designed stimuli compared to the variance for participants, simplifying the by-item random slope term is the first step.

\[
\text{lmer(DV} \sim \text{IV1} \ast \text{IV2} + \\
(1 + \text{IV1} \ast \text{IV2} \mid \text{participant}) + \\
(1 + \text{IV1} + \text{IV2} \mid \text{item}), \text{data})
\]

(2.2)

If this still fails then both random slope interaction terms can be simplified. The random slope can be further reduced so that it specifies a single dependent variable, or it can be removed completely if necessary.

\[
\text{lmer(DV} \sim \text{IV1} \ast \text{IV2} + \\
(1 + \text{IV1} + \text{IV2} \mid \text{participant}) + \\
(1 \mid \text{item}), \text{data})
\]

(2.3)

The models used in this dissertation retained the maximal random effects structure that would allow the model to converge. Any simplifications made to the default models are specified in the relevant sections. It should be noted that simplifying the random slope term affects the model’s interpretation of repeated measures. The only way to specify that a measure is repeated within-participants, for example, is to add the relevant fixed effect to the by-participant random slope. This simplification can increase the chance of Type I error. However, in the face of nonconvergence, this is the trade-off that must be made. In their evaluation of different model specifications, Barr et al. (2013) used a similar approach when
confronted with convergence issues. They also included the simplified models in the assessment of the maximal model performance from which they were derived. They argued that since they were “evaluating analysis strategies rather than particular model structures” (p. 266), this was a legitimate decision which also reflected common practices among researchers when dealing with nonconvergence.

Problems with convergence were occasionally encountered during the model comparison stage of the analyses in this dissertation. Removing a fixed effect in the comparison model (in order to calculate the p-value of the fixed effect) sometimes caused nonconvergence. Barr et al. (2013) also noted this phenomenon. Following their suggestion, the random effects structure of the comparison model was simplified according to the procedure described above. The model including the fixed effect in question was then re-fitted with the same random effects structure so that the likelihood ratio could be calculated. Simplifications made during model comparison are specified in the relevant sections.

One final point concerns whether a random slope for the looking control variable in the DAF experiments should be included in the models. While Barr et al. (2013) acknowledge that there has been limited formal testing of this issue, they suggest that it is not necessary to include a random slope for a control variable if there is no interaction with the fixed effects specified in the model. As such, the models in Chapter 3 do not include random effects for the looking variable.

### 2.6.3 Coding schemes

The default coding scheme used in this dissertation—and the default in the statistical software used—is treatment coding (also known as, dummy coding). This coding scheme compares the levels of the categorical independent variable to a reference level. In cases where collinearity (i.e. high correlations) among the fixed effects is an issue, sum coding can be used instead. This is a way of centering the variables, and compares the levels of the categorical independent variable to the grand mean. Note that collinearity is generally not an issue for experimental data since the independent variables are designed to be quite different and thus not likely to be correlated. Additionally, while coding choice affects the interpretation of the coefficients, it doesn’t affect how good a fit a model is. This means that model
Figure 2.2: Simulated data demonstrating homoscedasticity (left) and heteroscedasticity (right).

comparison—which is used in this dissertation to determine statistical significance—is not affected by coding choice.

2.6.4 Evaluating the models

Two assumptions of linear mixed effects model are that the residuals are normally distributed and homoscedastic. Homoscedasticity refers to the residuals having standard deviations that are constant across the range of fitted values. A simulation of homoscedastic residuals is shown in the lefthand plot of Figure 2.2. If the residuals do not have constant standard deviations across the range of fitted values, they are described as being heteroscedastic. A simulation of this is shown in the righthand plot of Figure 2.2.

For the linear mixed models constructed in this dissertation, residual plots were visually inspected to check for obvious deviations from homoscedasticity. If heteroscedasticity was observed a number of options were tested to improve the model, including fitting the data to a different distribution and transforming the data. Any such changes are specified in the relevant chapters.

For generalized linear mixed models using a Poisson distribution, an important assumption is that the variance of the errors increases with the mean (Baayen).
2008). The output of the modeling function in the glmmADMB package does not include an estimate of the ratio of variance and mean, so an estimate was calculated using a function proposed by developers of some of the commonly used (generalized) linear model packages for R. To meet the model assumptions, the ratio should be close to 1.

---

5 A description of this overdispersion function can be found at http://glmm.wikidot.com/faq.
Chapter 3

Visual feedback during a delayed auditory feedback task

3.1 Overview

This chapter reports on two experiments that examine the effect visual feedback has on speech produced with delayed auditory feedback. The first experiment randomized the presentation of two types of visual feedback—static and dynamic—while the second experiment separated the presentation of these different types of visual feedback and compared their effects to speech produced without visual feedback. Dynamic visual feedback affected both utterance duration and the number of speech errors in an utterance, but in different ways. Utterance durations, which were longer with delayed auditory feedback than normal auditory feedback, were further increased with dynamic visual in both experiments. Small differences in speech errors were found in the second experiment only. While the number of speech errors was reduced when dynamic visual feedback was paired with normal auditory feedback, this reduction was only significant when a measure of participants’ disruption by delayed auditory feedback was included in the statistical model; speech errors were reduced with dynamic visual feedback for participants who were minimally disrupted by delayed auditory feedback. These results suggest that it is the time-varying information in visual feedback which has the potential to elicit changes in speech production, and also that speakers need sustained exposure
to visual feedback in order for these changes to occur.

3.2 Introduction

3.2.1 Delayed auditory feedback

One method for investigating multimodal sensory processing is to create difficult speaking or listening conditions. For example, visual speech information improves perception accuracy to a greater extent when speech is presented in noise compared to clear speech conditions (Sumby and Pollack, 1954). For speech production, one way to create a difficult speaking task is to present speakers with delayed auditory feedback (DAF).

In a DAF task, speakers produce speech while their auditory feedback is played back to them with a delay. The length of the delay varies between experiments, ranging from very brief (e.g. 25 ms: Stuart et al., 2002) to very long (e.g. 500 ms: Howell et al., 1983). The most disruptive feedback delay is 180-200 ms, beyond which point the disruptive effects start to taper off (e.g. Black, 1951; Howell and Archer, 1984). It has been noted that in order for the delay to be most disruptive, the auditory feedback must be loud enough to mask feedback via bone conduction (e.g. Lee, 1950). Typical responses to DAF include reduced speech rate (or increased utterance duration), stuttering-like disfluencies, such as repetitions and prolongations, and increased intensity (e.g. Stuart et al., 2002; Yates, 1963). In a comprehensive investigation of the reported effects of DAF, Chesters et al. (2015) found significant effects on a range of measures: utterance duration, consonant duration, and vowel duration all increased, as was the total number of speech errors and the mean intensity. Mean pitch and pitch variation decreased, and there was a change to measures of rhythm. Disruption from DAF is amplified at fast speaking rates; Stuart et al. (2002) found significantly more speech errors during a reading passage spoken at a fast rate. Task differences have also been found; DAF induced speech rate decreases were greater for reading than conversation (Corey and Cuddapah, 2008).

The effect DAF has on speech kinematics has also been investigated. Starting from research findings that suggest that the response to a stimulus depends on
the point in the phase of activity to which it is applied, Zimmermann et al. (1988) looked at jaw movement relative to the onset and offset of the vocalic portions of DAF. With 100 ms DAF, the onset of DAF mostly occurred during the steady state voiced portion of syllable (defined as the period between the movement offset of the CV gesture and the movement onset of the VC gesture). With 200 ms DAF, the onset of DAF mostly occurred during the closing gesture of the syllable. Despite these different onset alignments, the offset for both DAF delays occurred mostly during the closed portion of the syllable (defined as the period between the movement offset of the VC gesture and the movement onset of the CV gesture). In other words, consonant closure durations were lengthened with 200 ms DAF relative to 100 ms DAF. Sasisekaran (2012) looked at lip aperture variability and lip movement duration in response to 200 ms DAF. As predicted, DAF resulted in greater values for both of these measures compared to unperturbed speech. This difference was also found in comparison to speech produced with gated feedback, suggesting that the effects are specific to DAF and not just any type of feedback perturbation.

Yates (1963) notes that there is a wide range of individual differences in response to DAF. Some of this can be attributed to gender differences; males tend to be more adversely affected (i.e. produce more speech errors) than females (Corey and Cuddapah, 2008). However, this isn’t the case for all measures of disruption: Chon et al. (2013) found that female participants exhibited a greater reduction in articulation rate than males, while Sasisekaran (2012) found that the duration of labial trajectories was longer for males than females. Yates (1963) also reports a possible correlation between verbal facility, or intelligibility, and the effect of DAF. Fabbro and Darò (1995) took up this question, comparing simultaneous interpreters, who have some experience with DAF, to a control group, and found that there was no difference in the number of speech errors between the normal and delayed conditions for the interpreters, but the control group made significantly more errors with DAF. The interpreters did have longer syllable durations with DAF, however, as did the control group. The authors suggest that speakers with high verbal facility may rely on auditory feedback less than speakers with lower verbal facility, and hence show a resistance to DAF.

Many researchers have noted that it is possible to counteract the effects of DAF by diverting attention away from the auditory feedback. Suggestions for how this
could be achieved include focusing on articulatory movements (Lee, 1950) or bone conducted feedback (Yates, 1963), or simply “not listening” Goldiamond et al. (1962). Lee (1950) also noted that participants attempted to slow their speech in order to avoid disfluencies.

This ability to counteract DAF effects has been explored in the context of adaptation studies. Following extended exposure to DAF, adaptation to the delay has been reported to occur to some degree, but generally not completely. There is also the potential for adaptation to be prevented by changing delay time or intensity (Yates, 1963). Katz and Lackner (1977) looked at these issues in more detail, with the goal of verifying that adaptation effects were not simply due to stabilization that occurs at the beginning of a task. Participants showed a reduction in speech errors and utterance duration, as well as aftereffects once the DAF was removed (increased speech rate). Additionally, it was observed that the degree of adaptation depended on the degree of disruption caused by the initial exposure to DAF; the more disrupted (in terms of both speech errors and duration) the greater the adaptation that followed. Further support for adaptation to DAF was found in later studies (e.g. Attanasio, 1987; Venkatagiri, 1980).

The participants in the Katz and Lackner (1977) study reported trying to use a variety of strategies to counteract the effects of DAF, ranging from changing their speech rhythm or speech rate, to focusing on their articulators. Attanasio (1987) explored the way that different strategies affect adaptation, comparing speakers who varied in their intrinsic sensitivity to somatosensory feedback (operationalized as the ability to discriminate between different shapes based on oral feedback), and by explicitly instructing half of the participants to focus on articulations. The results showed that the amount of adaptation was influenced by intrinsic oral awareness; speakers who had high sensitivity to oral somatosensory feedback adapted more than those with low sensitivity. However, receiving explicit instructions to focus on articulations did improve adaptation for speakers with low oral feedback sensitivity. The use of strategies led Katz and Lackner (1977) to propose that adaptation to DAF is an active process on the part of speakers, requiring “conscious strategic effort” (p. 482).
3.2.1.1 DAF with visual feedback

There is a small body of research which has investigated whether providing visual feedback would ameliorate the disruptive effects of DAF. These researchers hypothesized that, given the interaction of auditory and visual information in perception, a similar relationship may exist for production. DAF tasks were used since the visual signal would provide information that was synchronous with the actual speech output, and this could potentially counteract the effects of the asynchronous acoustic feedback. The results from this research have been mixed.

Tye-Murray (1986) was the first to investigate this. Participants read and memorized a sentence before producing it. During production they were presented with 200 ms DAF while simultaneously receiving visual feedback via a mirror. They were able to see from their neck to their mouth only, and they were specifically instructed to attend to the image of their mouth during the task. Utterance duration was measured, but no effect of visual feedback was found.

Jones and Striemer (2007) expanded this design; sentences were presented auditorily, which participants repeated with 180 ms DAF, and with the addition of noise to mask the real-time auditory feedback. The visual feedback condition was compared to a condition without visual feedback and one in which participants read the sentences. The number of speech errors per utterance was measured in addition to utterance duration. For the analysis, participants were split into two groups according to the number of speech errors they produced in the DAF condition with no visual feedback. Those who produced more errors were classified as high-disruption, and those who produced fewer errors were classified as low-disruption. While there was no effect of visual feedback overall for either of the variables measured, the low disruption group produced significantly fewer speech errors during production with DAF and visual feedback compared to the DAF reading condition.

In a pilot study to the Tye-Murray (1986) paper, the author had tested the effect of delayed visual feedback (2000 ms) on production. While overall there was no effect of the delay, two of the 13 participants produced errors that were clearly related to the visual feedback; for example, producing [st] instead of [soo] during visual alignment with [t]. This effect was explored more systematically by Chesters et al.
(2015), who tested both immediate and delayed visual feedback paired with 200 ms DAF. As with Jones and Striemer (2007), the sentences to be repeated were presented aurally, and masking noise was also presented throughout the experiment. Immediate visual feedback, which was provided by way of a mirror, did not reduce the effects of DAF; however there was an effect of visual feedback in the normal auditory feedback condition: the total duration of consonant segments was increased in this condition. Delayed visual feedback, which was provided by a delayed video stream and was either 200 ms, 400 ms, or 600 ms, had no significant effect on the speech measures when paired with normal auditory feedback but did in the DAF condition. Utterance duration, total number of speech errors, and total duration of vowel segments all had increased values, and there was also a change to rhythm. These differences held for all levels of the visual delay, except for the speech errors; speech errors increased with the 200 ms and 400 ms delays but not the 600 ms delay.

Overall then, results for the effect of visual feedback on production have been variable, both in terms of the direction of the effects and whether they are evident in combination with immediate or delayed auditory feedback.

3.2.1.2 A note on fluency and disfluency

Before moving forward, fluency and disfluency will be discussed in more detail, since the effects of DAF are discussed in these terms. While fluency can intuitively be thought of as “flowing” or “smooth” speech, it can be difficult to pinpoint specific features that contribute to this. In his review of these issues, Lickley (2014) discusses that part of the problem is the different levels/domains that fluency can be described at. For example, a listener’s perception of fluency versus measurable disturbances in the speech production, or, fluency in planning versus fluency in performance. Equally challenging is defining disfluency. As noted by (Lickley, 2014, p. 451-452), disfluencies are a normal part of speech: “While speakers may vary in the frequency of disfluencies in their speech, everyone is disfluent some of the time.” While there have been a range of proposals for what constitutes disfluency, recently there has been some consensus on disfluent features of normal speech; several features commonly identified are filled pauses, repetitions, substitutions,
insertions, and deletions (Lickley, 2014). Repetitions, substitutions, and insertions can involve words (either part or whole) or phrases. Prolongations are also commonly identified as a type of disfluency, especially in the context of experimentally induced disfluency (e.g. through DAF) and developmental stuttering (e.g. Corey and Cuddapah, 2008; Stuart et al., 2002).

While speech errors may seem an obvious example of disfluency, changes in speech rate, or utterance duration, may be less so. However, these measures are consistently found to be strong predictors of fluency for both non-native and native speech. For example, using a measure of mean length of syllables, Bosker et al. (2013) found this measure to be more highly negatively correlated with fluency ratings of non-native Dutch speech than other objective measures, such as number of pauses and number of corrections. Similar results were found with articulation rate for fluency ratings of French L2 speakers with certain tasks (Préfontaine et al., 2016). Using acoustic manipulations of speech rate (syllables per second including pauses) and articulation rate (syllables per second excluding pauses), Bosker et al. (2014) found that speed is weighted similarly for native and non-native speech when judging fluency. Temporal disfluency is a global increase in duration that goes beyond duration increases due to phonological processes such as phrase-final strengthening.

3.3 Experiment 1

As described above, speakers are known to actively use strategies to counteract the disruptive effects of DAF. Visual feedback that is provided during such a task could be used as one such strategy; for example, by focusing on the movement of the lips. In fact, most of the experiments using visual feedback have presented it in a way that makes this more likely. It is an open question as to whether visual feedback can still elicit changes when the opportunities for using it as a means to counteract DAF are minimized. This can be achieved by not directing attention to visual articulations and by changing the presentation order of stimuli.

Previous experiments have often presented visual feedback in a targeted manner. As described in Section 1.1.3, Snyder et al. (2009) looked at the effect visual feedback had on developmental stuttering. While they did observe improvements
in stuttering when visual feedback was presented, the effect was only found when participants actively attended to oral speech movements in the visual feedback signal. This observation was first made during the piloting stage of the study, and consequently explicit instructions were given in the main experiment for participants to attend to the motion of the articulators. In the context of non-disordered populations, the experiments in both Tye-Murray (1986) and Jones and Striemer (2007) did not give participants instructions to focus on the mouth; however, this was implicit in the fact that just the lower half of the face was presented as visual feedback. In contrast, Chesters et al. (2015) presented participants with a view of their whole face during visual feedback conditions, although they did not state what instructions were given for this condition. In the condition with no visual feedback, participants looked at a fixation point “at an equivalent position to their mouth when viewing their mirror image” (p. 876). This suggests that there was an expectation that participants would look at their mouth when presented with visual feedback, although it is unknown whether this was because they received instructions to do so.

Previous experiments have also presented visual feedback in consistent blocks. But varying the order of presentation of stimuli, either within or across experimental blocks, is one way to assess the robustness of multimodal speech processes. Taking the within-block case, multiple conditions can be presented in a single block instead of presenting one condition per block. In their investigation of the influence of auditory and haptic information on speech perception, Fowler and Dekle (1991) ran three experiments, two of which contrasted in the order of presentation of stimuli. In their first experiment, conditions were blocked such that audio only judgments of syllable identity were made in one block and audio-haptic judgments were made in a separate block. This procedure was changed in their third experiment so that trials from each condition were mixed in a single block; any one block comprised a random mix of audio only, haptic only, and audio-haptic judgments. They found a reduced effect of the felt syllable on the heard syllable in the third experiment compared to the first experiment, but the effect was nonetheless significant.

Order can also be manipulated across blocks. Varying the order of experimental blocks—where each block contains stimuli from a single conditions—is com-
monly used for counterbalancing. Rosenblum and Saldaña (1992) were interested in whether audiovisual perception was dependent on order of presentation. They varied the order of presentation of the audio, visual, and audiovisual conditions for their investigation of the phonetic similarity of congruous and incongruous audiovisual syllables. They found no effect of the order of presentation; that is, the visual influence on auditory percepts was robust to the type of stimuli that preceded it. In a later experiment investigating the importance of kinematics in visual speech information, Rosenblum and Saldaña (1996) found that the visual influence of static images on auditory perception was minimized when static images were presented after the dynamic images compared to before the dynamic images. In light of their earlier findings (Rosenblum and Saldaña, 1992), they suggested that this was evidence of a post-perceptual, or strategic, response to the static images, since the order of presentation shouldn’t affect audiovisual perception.

In the experiment presented in this chapter, both of these approaches are used to test for effects of visual feedback in a DAF task where it is less likely that active strategies will be used by speakers. The whole face was presented as feedback and participants were instructed to simply look at the monitor during the task. This presentation also capitalizes on the fact that speech relevant information is distributed across the entire face (Section 1.1.3). Instead of presenting each condition in a single block, the static and dynamic visual feedback conditions were randomized within a block, in a similar manner to the above experiment by Fowler and Dekle (1991). Static visual feedback was used as the baseline condition against which to compare the effects of dynamic visual feedback. Since the type of visual feedback being presented was randomized, a still image was used rather than a fixation point in order to avoid making the transition between the types of feedback jarring.

### 3.3.1 Hypothesis and predictions

As discussed in Section 1.1.3, the visual speech signal provides a range of speech information, from lower level articulatory information to higher level information about the temporal structure of speech. During speech production this signal is hypothesized to enhance speech output by providing information that is temporally compatible with other sources of feedback.
Given this hypothesis, visual feedback is predicted to counteract the disruptive effects of DAF by providing information that is synchronous with the non-delayed feedback (i.e. bone-conducted auditory feedback and somatosensory feedback). It is predicted to decrease both utterance duration and speech errors. As discussed in Section 3.2.1, increased utterance duration can be considered a type of disfluency, and is a common response to DAF. By reinforcing the correct speech timing, dynamic visual feedback is predicted to minimize this type of disfluency, as well speech errors.

3.3.2 Methods

3.3.2.1 Participants

26 students (mean age: 22.2 years (sd: 7.6); 7 males) recruited from The University of New Mexico participated in the experiment. Participants were compensated for their time with either a gift-card or extra credit toward their final grade. All participants self-reported having normal hearing and normal or corrected-to-normal vision, and all were native English speakers.

3.3.2.2 Stimuli

The stimuli consisted of 64 short sentences (mean length: 7.8 words (sd: 1); 8.8 syllables, (sd: 1); see Appendix A) taken from the Harvard sentences, a list of phonetically balanced sentences for use in speech tasks (Rothauser et al., 1969). The stimuli were recorded as per Section 2.3.

3.3.2.3 Procedure

The experiment, which required participants to repeat sentences in a four different feedback conditions, followed the procedure outlined in Section 2.4. A within-participants design with two independent variables was used. These variables manipulated the type of auditory and visual feedback participants received during speech production: auditory feedback was either normal or delayed (by 180 ms, to elicit the maximally disruptive effects of DAF), and visual feedback was either static or dynamic. Each participant was presented subsets of the stimuli in the
four conditions outlined in Table 3.1.

The stimuli were counterbalanced to conditions across participants using the procedure proposed by Durso (1984). This ensured that a particular stimulus appeared in each condition equally frequently, and any given participant responded to it only once. The conditions were organized into two types of mixed blocks. Normal auditory feedback blocks contained a randomized mix of NAF-Picture and NAF-Video conditions while delayed auditory feedback blocks contained a randomized mix of DAF-Picture and DAF-Video conditions. Each type of block was presented four times with each block containing a unique subset of the stimuli. The order of the blocks was counterbalanced across subjects. Each participant produced 64 sentences, yielding a total of 1664 sentences across all 26 participants. 15 sentences were discarded due to technical errors during the recording of the experiment.

White noise was presented throughout the experiment in an effort to mask participants’ real-time auditory feedback. During the presentation of the stimuli to be repeated the signal-to-noise ratio (SNR) was 10 dB. The SNR of the auditory feedback that participants received was not constant; while the noise level was held constant the signal level varied depending on how loudly the participant spoke.

### 3.3.2.4 Measures

Two commonly used measures of speech disruption in DAF paradigms are utterance duration (or speech rate) and total number of speech errors (Yates, 1963), both of which are used in this experiment.

<table>
<thead>
<tr>
<th>auditory feedback</th>
<th>visual feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>static</td>
</tr>
<tr>
<td></td>
<td>NAF-Picture</td>
</tr>
<tr>
<td></td>
<td>NAF-Video</td>
</tr>
<tr>
<td>delayed</td>
<td>dynamic</td>
</tr>
<tr>
<td></td>
<td>DAF-Picture</td>
</tr>
<tr>
<td></td>
<td>DAF-Video</td>
</tr>
</tbody>
</table>

**Table 3.1:** Conditions presented to each participant.
In line with the three previous studies which investigated visual feedback effects during a DAF task (Chesters et al., 2015; Jones and Striemer, 2007; Tye-Murray, 1986), utterance duration was measured. To maintain consistency in segmentation, a set of criteria were established for identifying the beginning and end of an utterance, based on an item by item comparison of seven of the participants. As an example, utterances that ended with a stop might or might not have an audible release. In the case of the audible release, the end of the utterance was marked at the end of the release once there was no longer any evidence of F2 in the burst. If there was no audible release, the utterance was marked as ending at the end of voicing.

Speech errors were coded following the disfluency transcription conventions proposed in Brugos and Shattuck-Hufnagel (2012). Table 3.2 presents a summary of the coding scheme. These conventions build on a number of proposals for annotating disfluent speech and cover all of the disfluent phenomena produced by participants in this experiment. An additional benefit of this coding scheme is that it was designed to be used in Praat (Boersma and Weenink, 2009), which was the program used for duration analysis. Some of the error codes were adapted to the particulars of this experiment. For example, since the task involved repeating sentences rather than producing unscripted utterances, there was the potential for words to be deleted during the repetition. These missing words were coded as an “e” error and were marked between the two words where they should have occurred. This type of error involved missing content words; for example, if clean was deleted during the repetition of The doorknob was made of bright clean brass. Any cases that could be considered a natural reduction were not marked as an error; for example, if You cannot brew tea in a cold pot was repeated as You can’t brew tea in a cold pot, the reduced modal was not coded as an error. If a speech error occurred in the word immediately following the missing word, the missing word error was coded with that speech error. An example of a complex disfluent event (i.e. one with multiple speech errors) is provided in Figure 3.1.

Some words from the stimuli sentences were consistently produced as a phonetically similar word by a large number of participants. Examples include producing “green” instead of “clean”, “meals” instead of “mules”, and “pushed” instead of “plush.” All of these cases resulted in a semantically plausible sentence. It was
<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>prolongation</td>
<td>pr</td>
<td>abnormal and/or incongruous prolongation of a segment within a word</td>
</tr>
<tr>
<td>disfluent pause</td>
<td>ps, psw</td>
<td>abnormal and/or incongruous pause between (ps) or within (psw) words</td>
</tr>
<tr>
<td>silence</td>
<td>s</td>
<td>end of a silence (whether disfluent-sounding or not)</td>
</tr>
<tr>
<td>filler</td>
<td>f</td>
<td>filled pause, filler words or segments (e.g. um, huh, mm)</td>
</tr>
<tr>
<td>error</td>
<td>e</td>
<td>mispronunciation or wrong word</td>
</tr>
<tr>
<td>cut</td>
<td>c</td>
<td>a partially completed word</td>
</tr>
<tr>
<td>restart word</td>
<td>rs</td>
<td>restarting of a segment, syllable, word, after a word has been cut off</td>
</tr>
<tr>
<td>restart phrase</td>
<td>%r</td>
<td>start of a new phrase after a previous phrase was not finished</td>
</tr>
</tbody>
</table>

**Table 3.2:** Disfluency transcription conventions from Brugos and Shattuck-Hufnagel (2012).

decided not to categorize these as production errors, since they were likely due to misperception when initially listening to the stimuli (which were embedded in noise). A list of these accepted substitutions is included in Appendix A.

The most difficult type of disfluency to mark was prolongation in the DAF conditions, as these conditions were produced with an overall slower speech rate. In keeping with Brugos and Shattuck-Hufnagel’s definition of prolongation (“abnormal and/or incongruous prolongation”), prolongation was judged relative to the utterance in which it occurred in order to decide whether it was abnormal or incongruent. This allowed normal prosodic lengthening to be distinguished from disfluent prolongation. It also helped to determine if a potential prolongation was due simply to an overall slowed speech rate, or if it was an additional source of lengthening and thus incongruous with the rest of the utterance. Thus, what counted as
Soap can wash most dirt away

Figure 3.1: An example of a complex disfluent event (error code: e.ps). The speaker makes a pronunciation error (*watch* instead of *wash*) which is immediately followed by a pause.

an instance of prolongation varied from item to item and from participant to participant. Additionally, both part- and whole-word prolongations were coded. An example of a part-word prolongation is given in Figure 3.2.

During error coding, items were discarded if they included laughter, were incomplete, or there was a major recall error. 37 items were discarded during the error coding process. The final number of items in the analysis was 1612.

Error coders were blind to the conditions being coded. It should be noted that it was usually obvious which auditory feedback condition a participant was performing, due to the striking effect DAF has on speech production. However, the critical comparisons for this experiment were the visual feedback conditions, and coders were not able to identify this aspect of the experimental manipulation.

---

1An example of a major recall error: For the stimulus *The wreck occurred by the bank on Main Street*, one participant instead produced *The wreck recorded at the red street*. This item was discarded.
The fur of cats goes by many names.

Figure 3.2: An example of a part-word prolongation. The labial nasal in *many* is abnormally prolonged, with a duration of approx. 240 ms.

### 3.3.2.5 Inter-rater agreement

A second coder coded 13% of the utterances. These utterances were randomly selected from the data set, with the restriction that the selection include two sentences per condition per subject. The second coder was given detailed instructions about implementing the coding scheme and also went through a set of examples with the main coder. Inter-rater word-by-word agreement was Cohen’s *kappa* 0.662 (p < 0.001). Cohen’s *kappa* values between 0.61 and 0.80 represent substantial agreement beyond chance (Landis and Koch, 1977). Table 3.3 shows the cross-tabulation of errors categories from each coder.

### 3.3.2.6 Video coding

As discussed in Section 2.4, the video recordings were coded for whether or not each participant was looking at the monitor while they produced each utterance. 452 of the 1612 items were produced while the participant was not looking at the monitor. This looking variable was included in the statistical analyses below.
<table>
<thead>
<tr>
<th></th>
<th>Main Coder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ø</td>
<td>1162</td>
</tr>
<tr>
<td>e</td>
<td>27</td>
</tr>
<tr>
<td>ps</td>
<td>1</td>
</tr>
<tr>
<td>pr</td>
<td>13</td>
</tr>
<tr>
<td>rs</td>
<td>5</td>
</tr>
<tr>
<td>c</td>
<td>7</td>
</tr>
</tbody>
</table>

**Table 3.3:** Cross-tabulation of error categories from each coder. (Ø = no error, e = mispronunciation or word error, ps = pause, pr = prolongation, rs = restart, c = cut)

<table>
<thead>
<tr>
<th></th>
<th>Picture</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utterance duration (ms)</td>
<td>NAF 2235 (66)</td>
<td>NAF 2253 (66)</td>
</tr>
<tr>
<td></td>
<td>DAF 3479 (76)</td>
<td>DAF 3539 (66)</td>
</tr>
<tr>
<td>No. speech errors per utterance</td>
<td>NAF 0.36 (0.06)</td>
<td>NAF 0.39 (0.06)</td>
</tr>
<tr>
<td></td>
<td>DAF 1.38 (0.07)</td>
<td>DAF 1.41 (0.08)</td>
</tr>
</tbody>
</table>

**Table 3.4:** Means and standard errors of the means of the experimental measures for each condition.

### 3.3.3 Results

The means and standard errors of the means for utterance duration and total speech errors are shown in Table 3.4. The statistical results for each measure are discussed below.
Figure 3.3: Distributions of utterance duration for each condition. The lower and upper hinges of the boxplot represent the first and third quartiles, respectively. The middle line represents the median.

### 3.3.3.1 Utterance duration

A summary of the utterance duration results is shown in [Figure 3.3](#). The model construction, criticism, and comparison steps are outlined below.

A linear mixed effects analysis was performed using the general formula outlined in [Section 2.6](#). The two independent variables were Type of Auditory Feedback and Type of Visual Feedback, and the Looking variable was also added. The maximal random effects structure was specified. Visual inspection of the residual plots of this model revealed heteroscedasticity.

The distribution of the duration measures revealed right-skewing suggestive of a log-normal distribution, so the values were log-transformed and the models were re-fit. During the model comparison process the random effects structure had to be simplified for the model that contained only the auditory feedback factor in order
<table>
<thead>
<tr>
<th></th>
<th>coefficient</th>
<th>(std. error)</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>7.68</td>
<td>(0.03)</td>
<td>291.89</td>
</tr>
<tr>
<td>auditory(DAF)</td>
<td>0.42</td>
<td>(0.03)</td>
<td>15.22</td>
</tr>
<tr>
<td>visual(video)</td>
<td>0.02</td>
<td>(0.01)</td>
<td>2.50</td>
</tr>
<tr>
<td>looking(no)</td>
<td>0.03</td>
<td>(0.01)</td>
<td>3.21</td>
</tr>
</tbody>
</table>

Table 3.5: Fixed effects for the model of (log) utterance duration. The significance of the auditory feedback and visual feedback factors were confirmed with model comparison.

for the model to converge. This model and its comparison model had simplified by-item random slopes, allowing variation for the auditory feedback factor only. The random effect for participants retained the maximal specification. Visual inspection of the residual plots for these models revealed no obvious deviations from homoscedasticity.

Model comparison revealed a significant effect of auditory feedback ($\chi^2(1) = 59.133$, $p < .001$): utterance duration increased in the DAF conditions compared to the NAF conditions. There was also a significant effect of visual feedback ($\chi^2(1) = 6.3754$, $p = .01157$), with increased utterance duration when visual feedback was the video compared to the picture. Model comparison failed to find an effect of the interaction between auditory and visual feedback. Table 3.5 shows the fixed effects coefficients, standard errors, and t-values for the model with the significant main effects.

As seen in Table 3.5 the Looking control variable was a significant factor in the model. Utterance duration increased when participants were looking away from the monitor compared to when they were looking at the monitor.

### 3.3.3.2 Speech errors

The counts for each type of error are shown in Figure 3.4. Mispronunciations and word errors, which included incorrect or missing words and word inversions (e.g. clean bright brass instead of bright clean brass), were the most common type of
error. However, for the DAF-Picture condition, errors, restarts, prolongations, and pauses had very similar counts. Other than mispronunciations and word errors, the other types of errors were much more common in the DAF conditions than the NAF conditions. This was especially the case for restarts and prolongations, two stuttering-like disfluencies commonly produced under DAF conditions.

A summary of the total number of speech errors per utterance is shown in Figure 3.5. A generalized linear mixed effects analysis (Poisson distribution) was performed using the general formula outlined in Section 2.6. The two independent variables were Type of Auditory Feedback and Type of Visual Feedback, and the Looking control variable was also added. The maximal random effects structure was specified. The models also accounted for zero-inflation in the data.

During the model comparison process the random effects structure had to be simplified for the model that contained only the visual feedback factor in order
Figure 3.5: Density plots of speech errors for each condition. The vertical dashed lines represent the mean number of speech errors per utterance for each condition.

for the model to converge. This model and its comparison model had a maximal by-participant random slope but an intercept only for item. The ratio of the error variance to mean was 1.03 for the full model, suggesting that there were no concerns of overdispersion.

Model comparison revealed a significant effect of auditory feedback ($D = 59.36$, $p < .001$).\textsuperscript{2} As expected, speech errors increased in the DAF conditions compared to the NAF conditions. Model comparison failed to find an effect of visual feedback or an interaction between the independent variables. Table 3.5 shows the fixed effects coefficients, standard errors, and $z$-values for the model with main effects but no interaction.

As seen in Table 3.6, the Looking control variable was once again significant.

\textsuperscript{2}The likelihood ratio test for generalized linear models uses deviance ($D$) as a measure of model fit. [Zuur et al. (2009)] describe it as a maximum likelihood equivalent of the sum of squares of residuals.
Table 3.6: Fixed effects for the model of speech errors. The significance of the auditory feedback factor was confirmed with model comparison. (NB: Since the model used a Poisson distribution, the coefficients are in expected log counts.)

<table>
<thead>
<tr>
<th>Effect</th>
<th>Coefficient (Std. Error) z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.25 (0.12) -10.57</td>
</tr>
<tr>
<td>auditory(DAF)</td>
<td>1.32 (0.09) 14.24</td>
</tr>
<tr>
<td>visual(video)</td>
<td>0.08 (0.06) 1.19</td>
</tr>
<tr>
<td>looking(no)</td>
<td>0.25 (0.07) 3.54</td>
</tr>
</tbody>
</table>

The total number of speech errors increased when participants were looking away from the monitor compared to when they were looking at the monitor.

3.3.3.3 Post-hoc analysis: DAF disruption

The planned statistical analysis failed to find an effect of visual feedback on the number of speech errors. The following analysis looks at the speech error results in more detail. It should be noted that this was not one of the planned comparisons for this experiment, but is presented to gain more insight into possible effects of visual feedback.

As discussed in Section 3.2.1, Jones and Striemer (2007) tested whether there was a differential effect of visual feedback based on the degree to which a participant was affected by DAF. They split participants into two groups based on the number of speech errors produced in the baseline DAF condition (i.e. no visual feedback). Participants who made fewer errors were categorized as “low-disruption” and participants who made more errors were categorized as “high-disruption.” They found that participants in the low disruption group produced few errors in the DAF condition with visual feedback compared to the DAF condition in which stimuli were read off the monitor.

This interaction was tested in the present data set. Instead of splitting participants into two groups based on their mean number of errors in the DAF-Picture
Table 3.7: Fixed effects for the model of speech errors with the significant effect of DAF disruption. (NB: Since the model used a Poisson distribution, the coefficients are in expected log counts.)

<table>
<thead>
<tr>
<th>Condition</th>
<th>Coefficient</th>
<th>(Std. Error)</th>
<th>z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.06</td>
<td>(0.19)</td>
<td>-10.93</td>
</tr>
<tr>
<td>auditory(DAF)</td>
<td>1.47</td>
<td>(0.13)</td>
<td>11.48</td>
</tr>
<tr>
<td>visual(video)</td>
<td>0.08</td>
<td>(0.17)</td>
<td>0.47</td>
</tr>
<tr>
<td>disruption</td>
<td>0.59</td>
<td>(0.11)</td>
<td>5.38</td>
</tr>
<tr>
<td>auditory(DAF):visual(video)</td>
<td>-0.06</td>
<td>(0.17)</td>
<td>-0.36</td>
</tr>
</tbody>
</table>

Condition, the mean values were added to the model as a continuous variable, which will be referred to as “disruption.” This variable is interpreted as a measure of a speaker’s verbal proficiency; the less disrupted a speaker is by DAF the more proficient they are (Chon et al., 2013).

A generalized linear mixed effects analysis (Poisson distribution) was performed, with Type of Auditory Feedback, Type of Visual Feedback, and Disruption as independent variables. The random effects structure was simplified to achieve convergence; the by-participant random slope remained maximal but the random effect for item contained an intercept only.

Model comparison revealed a significant effect of disruption ($D = 13.92, p < .001$), with the number of speech errors increasing across the board for participants who were more disrupted by DAF. Model comparison failed to find an interaction between disruption and the feedback variables. Table 3.7 shows the fixed effects coefficients, standard errors, and z-values for the model with the significant effect of disruption, and this effect is visualized in Figure 3.6.

3.3.4 Discussion

The results of this analysis showed an effect of dynamic visual feedback on utterance duration; however, it was in the opposite direction to the prediction. For both NAF and DAF conditions, utterance duration increased with dynamic visual
feedback compared to static visual feedback. The analysis failed to find an effect of dynamic visual feedback on speech errors. A post-hoc analysis also failed to find any interaction with degree of DAF disruption.

Visual feedback was predicted to decrease utterance duration, but instead the results showed an increase in duration for the DAF-Video condition. This may be related to reports of increased segmental durations in two similar contexts to the present experiment. As discussed in Section 3.2.1, Zimmermann et al. (1988)
looked at jaw movement in DAF tasks, comparing two auditory feedback delays: 100 ms and 200 ms. Despite the fact that the onset of DAF aligned with different articulatory-defined portions of the syllable, the offset of DAF most often occurred during the closed portion of the syllable for both delays. In order for this offset alignment to occur, the closed gesture must be sustained for a longer period of time for the 200 ms DAF relative to 100 ms DAF. Recall that in the present experiment, DAF was 180 ms, so closer to the 200 ms delay than the 100 ms delay in Zimmermann et al. (1988). This DAF offset alignment pattern was primarily found when the following syllable had a bilabial or labiodental onset, two places of articulation that are visually distinct. It is possible that these sustained closed portions of the syllable were reinforced by dynamic visual feedback, thus lengthening them further still. Prolongation is a commonly reported disfluency under DAF conditions, and this was also observed in the present experiment (Figure 3.4). Chesters et al. (2015) corroborated the prolongation results from subjective speech error coding with measures of consonant duration; they found increased consonant duration with DAF compared to NAF. Consonant duration was also increased with visual feedback, although this effect was only significant when auditory feedback was normal. If a similar effect occurred in the present experiment, this could account for the increased utterance duration in the NAF-Video condition.

Jones and Striemer (2007) reported a reduction in speech errors for a subset of their participants when visual feedback was presented during a DAF task, but this was not found in the present experiment. One of the biggest differences between these two experiments is the stimulus presentation order. As described in Section 3.3, static and dynamic visual feedback were randomized within a block in order to make it less likely that participants would develop a strategy for using the visual feedback to counteract the disruptive effects of DAF. But it may be the case that sustained exposure to visual feedback over the course of a block is necessary for visual feedback effects to be observed. The next experiment uses a similar design to the present experiment, but presents one condition per block (as in Chesters et al., 2015; Jones and Striemer, 2007; Tye-Murray, 1986). A condition without any visual feedback is also added, against which to compare the effects of each type of visual feedback (static and dynamic).
3.4 Experiment 2

In the previous experiment, static visual feedback was used as a baseline against which to compare the effects of dynamic visual feedback. Experiment 2 involves the same task as Experiment 1 (i.e. participants repeated sentences with or without DAF while being presented with different types of visual feedback) and examines whether there are different responses to these two types of visual signals by comparing them to speech production without visual feedback.

While speech is a time-varying phenomenon, it is possible to capture certain linguistically relevant distinctions from static visual forms (Fromkin, 1964). A variety of studies with fluent speakers suggest there are differences in the way static and dynamic visual speech information is processed in perception tasks. Several studies have documented improvements in perception tasks when dynamic—but not static—visual speech is presented with congruous auditory stimuli. For example, dynamic information improved response times when identifying speech from non-speech (Kim and Davis, 2014) and in a syllable identification task (Gilbert et al., 2012). The N1 component of the event-related brain potential is sensitive to auditory speech and can be modulated by visual speech. EEG recordings reveal decreases in latency and amplitude of N1 in response to dynamic visual information compared to other visual signals (Gilbert et al., 2012). The latency improvement was seen for any kind of motion (chewing and speech) whereas the amplitude decrease was specific to speech motion, which may be indicative of an alerting mechanism; any kind of facial motion enhances the speed of the auditory neural response, with meaningfulness of the visual information (i.e. speech or non-speech) subsequently reflected in the magnitude of the response.

Such studies suggest no role for static visual information; however, there is evidence that this information can influence perception. Using a McGurk paradigm, still images of a face affected responses to auditory stimuli, increasing the percentage of visually influenced percepts (Rosenblum and Saldaña, 1996). The authors suggested that this effect on auditory perception was due to post-perceptual processing (i.e. strategic responding) based on the results of a follow-up experiment; they found that static images had a reduced effect on responses to auditory stimuli when they were preceded by conditions which had dynamic images. Earlier
work from the authors had shown that the order of presentation of McGurk-type stimuli should not affect the influence visual information has on auditory perception (Rosenblum and Saldaña, 1992). Irwin et al. (2006) also contrasted static and dynamic images using a McGurk paradigm, with conflicting results. They used brief visual stimuli (approx. 100 ms). For the static condition the video frame that most clearly showed the consonant place of articulation was repeated three times, while for the dynamic condition three consecutive frames that spanned the consonant closure and release were presented. When the temporal alignment between the audio signal and the different types of visual signals was controlled for, there was no difference in the effect of the different types of visual signals. However, in a visual-only identification task, the brief static images were more accurately identified than the brief dynamic images. This result is unexpected, but may be due to the fact that a 100 ms image composed of a single frame is visually ‘clearer’ than a 100 ms image composed of three frames. Thus the result may say more about the effects of visual clarity than it does about static versus dynamic perception. Additionally, in an experiment looking at the effect of external visual signals on stuttering inhibition, static and dynamic visual information were equally effective in reducing stuttering when combined with a pure tone (Guntupalli et al., 2011).

Neuroimaging suggest that both static and dynamic visual speech images are relevant to auditory cortical processing, although with differences in the degree of activation. In a phoneme detection task, Calvert and Campbell (2003) presented normal hearing participants with either dynamic visual images of syllables (e.g. [vu], [im]) or static images from the maximal closed or open portion of these syllables (e.g. [v], [u]). fMRI results showed that, overall, similar cortical regions were activated for both types of visual speech signals, but dynamic visual speech activated these areas to a greater degree. However, there were some differences: dynamic images activated visual motion areas, auditory cortex, and the left superior temporal sulcus to a greater extent (but static images did activate these areas to some extent; this could be seen when compared to the baseline condition); static images activated the ventral premotor cortex and intraparietal cortex to a greater degree. So, while static images can influence auditory areas, this influence is enhanced by the matching temporal structure between dynamic visual speech and auditory visual speech.
Studies of neurophysiologically impaired patients also support distinct processing of static and dynamic visual speech information, complementing the behavioral and neuroimaging results from normal populations. Campbell (1992) presented evidence of differences in the processing of static and dynamic speech information. Similarly, a subject with profound visual form agnosia who performed on par with control subjects in audiovisual, auditory only, and visual only vowel identification tasks, dropped to chance performance when static images were presented (Munhall et al., 2002). Other patients with visual agnosia have exhibited similar differences in static and dynamic visual processing, although the responses to audiovisual stimuli suggest that the audiovisual integration may be compromised (de Gelder et al., 1998).

These studies show impairment to static visual processing while sparing dynamic visual processing. But there is a double dissociation, as the opposite pattern is also attested. In a well-documented case of ‘motion-blindness’, a patient with a lesion in the V5 region of the occipital lobe was able to identify speech forms from static images of the face but was unable to repeat multisyllabic forms presented visually and was unable to integrate visual information when presented with incongruous audiovisual syllables (Campbell et al., 1997). The results from behavioral and neuroimaging experiments, as well as lesion studies, thus show the importance of both static form and dynamic timing properties for visual speech processing.

3.4.1 Hypothesis and prediction

While the time-varying properties of the visual feedback are considered to be important for establishing temporal compatibility with other feedback signals, there may still be an enhancing effect of static visual feedback, albeit weaker than dynamic visual feedback. This is in line with the finding from Calvert and Campbell (2003) that static visual speech activated similar cortical regions to dynamic visual speech but to a lesser degree. Experiment 2 compares utterances produced with these two forms of visual feedback, and also makes a comparison to utterances produced without visual feedback. In Experiment 2 the different types of visual feedback are presented in separate blocks; this consistent exposure to a single type of visual feedback is similar to the experimental procedures used in previous ex-
periments which pair visual feedback and DAF (Chesters et al., 2015; Jones and Striemer, 2007; Tye-Murray, 1986). Presenting the visual feedback in this way, as opposed to the randomized presentation used in Experiment 1, may make it more likely to see effects of visual feedback on speech production.

As in Experiment 1, visual feedback is predicted to counteract the disruptive effects of DAF by providing information that is synchronous with the non-delayed feedback. Dynamic visual feedback—a rich source of time-varying speech information—is predicted to have the greatest impact on speech production. In terms of utterance duration, dynamic visual feedback could have a fluency-enhancing effect and thus decrease duration; although given the results of Experiment 1, it is also possible that dynamic visual feedback could increase utterance duration. In terms of speech errors, dynamic visual feedback is predicted to decrease the number of speech errors. Static visual feedback is predicted to have qualitatively similar effects but with a reduction in the magnitude of the effects.

### 3.4.2 Methods

#### 3.4.2.1 Participants

Thirty students (mean age: 21.2 years (sd: 4.5); 6 males) recruited from The University of New Mexico’s Linguistics 101 course participated in the experiment. Participants were compensated for their time with extra credit toward their final grade. All participants self-reported having normal hearing and normal or corrected-to-normal vision, and all were native English speakers.

#### 3.4.2.2 Stimuli

The stimuli consisted of 60 short sentences (mean length: 7.8 words (sd: 1); 8.7 syllables, (sd: 1); see Appendix A) taken from the Harvard sentences, a list of phonetically balanced sentences for use in speech tasks (Rothauser et al., 1969). The stimuli were recorded as per Section 2.3.
Table 3.8: Conditions presented to each participant.

<table>
<thead>
<tr>
<th>auditory feedback</th>
<th>visual feedback</th>
<th>absent</th>
<th>static</th>
<th>dynamic</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>NAF-NoVideo</td>
<td>NAF-Picture</td>
<td>NAF-Video</td>
<td></td>
</tr>
<tr>
<td>delayed</td>
<td>DAF-NoVideo</td>
<td>DAF-Picture</td>
<td>DAF-Video</td>
<td></td>
</tr>
</tbody>
</table>

3.4.2.3 Procedure

The experiment, which required participants to repeat sentences in six different feedback conditions, followed the procedure outlined in Section 2.4. A within-participants design with two independent variables was used. These variables manipulated the type of auditory and visual feedback participants received during speech production: auditory feedback was either normal or delayed (by 180 ms), and visual feedback was either absent, static, or dynamic. This resulted in six conditions shown in Table 3.8. Each participant produced subsets of the stimuli in the six conditions.

The stimuli were grouped into blocks of 10 sentences, with each block comprised of a single condition. Condition order and stimuli assignment were simultaneously counterbalanced using the procedure proposed by Zeelenberg and Pecher (2014). This controlled sequential and ordinal effects of condition ordering and ensured a particular stimulus appeared in each condition equally frequently, and any given participant responded to it only once. Each participant produced 60 sentences, yielding a total of 1800 sentences across all 30 participants. 11 sentences were discarded due to technical errors during the recording of the experiment.

White noise was presented throughout the experiment in an effort to mask participants’ real-time auditory feedback. During the presentation of the stimuli to be repeated, the signal-to-noise ratio (SNR) was 10 dB. The SNR of the auditory feedback that participants received was not constant; while the noise level was held constant, the signal level varied depending on how loudly the participant spoke.
### Table 3.9: Cross-tabulation of error categories from each coder. (Ø = no error, e = mispronunciation or word error, ps = pause, pr = prolongation, rs = restart, c = cut)

<table>
<thead>
<tr>
<th></th>
<th>Ø</th>
<th>e</th>
<th>ps</th>
<th>pr</th>
<th>rs</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Second Coder</td>
<td>1270</td>
<td>18</td>
<td>6</td>
<td>8</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>e</td>
<td>5</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ps</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>pr</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rs</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

3.4.2.4 Measures

The measures were the same as those described in Section 3.3.2.4 for Experiment 1; namely, utterance duration and speech errors. 24 items were discarded based on the error coding criteria outlined in Section 3.3.2.4. The final number of items in the analysis was 1765.

3.4.2.5 Inter-rater agreement

A second coder coded 10% of the utterances. These utterances were randomly selected from the data set, with the restriction that the selection include one sentence per condition per participant. Inter-rater word-by-word agreement was Cohen’s kappa 0.695 (p < 0.001). Cohen’s kappa values between 0.61 and 0.80 represent substantial agreement beyond chance (Landis and Koch, 1977). Table 3.9 shows the cross-tabulation of errors categories from each coder.

3.4.2.6 Video coding

As discussed in Section 2.4, the video recordings were coded for whether or not each participant was looking at the monitor while they produced each utterance.
333 of the 1765 items were produced while the participant was not looking at the monitor. This looking variable was included in the statistical analyses below.

### 3.4.3 Results

The means and standard errors of the means for utterance duration and total speech errors are shown in Table 3.10. The statistical results for each measure are discussed below.

#### 3.4.3.1 Utterance duration

A summary of the utterance duration results is shown in Figure 3.7. The statistical analysis involved a number of iterations of model construction before an adequate fit was found that also met the assumptions for model validity. These steps are outlined below.

A linear mixed effects analysis was performed using the general formula outlined in Section 2.6. The two independent variables were Type of Auditory Feedback and Type of Visual Feedback, and the Looking control variable was also added. The maximal random effects structure was specified. Visual inspection of the residual plots of these models revealed heteroscedasticity.

To address this the values were log-transformed, as per Experiment 1. The model failed to converge with the maximal random effects structure. The by-item random slope was simplified so that it no longer had an interaction term for the

<table>
<thead>
<tr>
<th>Utterance duration (ms)</th>
<th>No Video</th>
<th>Picture</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAF 2188 (62)</td>
<td>2172 (53)</td>
<td>2125 (56)</td>
<td></td>
</tr>
<tr>
<td>DAF 3084 (90)</td>
<td>3084 (51)</td>
<td>3194 (78)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No. speech errors per utterance</th>
<th>NAF</th>
<th>DAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.39 (0.06)</td>
<td>0.98 (0.07)</td>
<td></td>
</tr>
<tr>
<td>0.41 (0.06)</td>
<td>1.00 (0.06)</td>
<td></td>
</tr>
<tr>
<td>0.29 (0.05)</td>
<td>0.97 (0.07)</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.10:** Means and standard errors of the means of the experimental measures for each condition.
effect of type of auditory and visual feedback. Visual inspection of the residual plots revealed no obvious deviations from homoscedasticity.

During the model comparison process the random effects structure had to be further simplified for the model that contained only the visual feedback factor. This model, and its comparison model, had an intercept only for the random effect of item.

Model comparison revealed a significant effect of auditory feedback ($\chi^2(1) = 58.934, p < .001$); utterance duration increased in the DAF conditions compared the NAF conditions. The interaction between the independent variables was also significant ($\chi^2(1) = 8.9253, p = .01153$). Table 3.11 shows the fixed effects coefficients, standard errors, and t-values for the model with the significant interaction. The contrasts of interest for the interaction were obtained using multcomp
Table 3.11: Fixed effects for the model of (log) utterance duration with a significant interaction of the predictors.

(Hothorn et al., 2008). Post-hoc comparisons between the levels of the visual feedback variable were made for each level of the auditory feedback variable. For the DAF conditions there was a significant increase in utterance duration with dynamic visual feedback compared to no visual feedback (z = 2.956, p < .01 (adjusted p-value, single step method)) and compared to static visual feedback (z = 2.561, p = .02823 (adjusted p-value, single step method)). However, there was no significant difference in utterance duration between static visual feedback and no visual feedback, nor were there any significant differences between the types of visual feedback in the NAF conditions.

### 3.4.3.2 Speech errors

The counts for each type of error are shown in Figure 3.8. Mispronunciations and word errors accounted for most of the errors across the conditions, and they accounted for a larger proportion of the total errors than in Experiment 1. As with the previous experiment, the stuttering-like disfluencies (i.e. restarts and prolongations) were more common in the DAF conditions than the NAF conditions.

A summary of the total number of speech errors per utterance is shown in Figure 3.9. Since the dependent variable was a count variable, a generalized lin-
ear mixed effects analysis (Poisson distribution) was performed using the general formula outlined in Section 2.6. The two independent variables were Type of Auditory Feedback and Type of Visual Feedback, and the Looking control variable was also added. The maximal random effects structure was specified. The models also accounted for zero-inflation in the data.

The full model failed to converge with the maximal random effects structure so it was simplified by removing the interaction term from the by-item random slope. The ratio of the error variance to mean was approximately 1 for this new model, suggesting that there were no concerns of overdispersion.

During the model comparison process the random effects structure had to be further simplified for the model that contained only the auditory feedback factor. This model, and its comparison model, had a by-participant random slope for the effect of the auditory feedback factor, and an intercept only for the random effect.
Model comparison revealed a significant main effect of auditory feedback ($D = 59.48$, $p < .001$); speech errors increased in the DAF conditions compared to the NAF conditions. Model comparison failed to find an effect of visual feedback or an interaction between the independent variables. Table 3.12 shows the fixed effects coefficients, standard errors, and z-values for the model with the significant auditory feedback effect.

As can be seen in Table 3.12, the Looking control variable was a significant factor in the model. The total number of speech errors increased when participants were looking away from the monitor compared to when they were looking at the monitor.
Table 3.12: Fixed effects for the model of speech errors with the significant auditory feedback effect. (NB: Since the model used a Poisson distribution, the coefficients are in expected log counts.)

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>(Std. Error)</th>
<th>z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.02</td>
<td>(0.12)</td>
<td>-8.55</td>
</tr>
<tr>
<td>auditory(DAF)</td>
<td>0.95</td>
<td>(0.08)</td>
<td>12.26</td>
</tr>
<tr>
<td>visual(picture)</td>
<td>0.03</td>
<td>(0.07)</td>
<td>0.35</td>
</tr>
<tr>
<td>visual(video)</td>
<td>-0.07</td>
<td>(0.08)</td>
<td>-0.89</td>
</tr>
<tr>
<td>looking(no)</td>
<td>0.19</td>
<td>(0.08)</td>
<td>2.28</td>
</tr>
</tbody>
</table>

3.4.3.3 DAF disruption

As in Experiment 1, the speech error results were explored in more detail by testing whether changes in speech output in response to visual feedback were sensitive to the degree of DAF disruption a participant experienced.

Given that the Looking control variable was significant in the speech error model, it was decided to exclude those items from the analysis in which the participant was not looking at the monitor. This left 1432 sentences in the analysis. The disruption variable was calculated over this subset. One participant did not look at the monitor at all during the DAF-NoVideo condition, so all data points from that participant were excluded. This further reduced the number of sentences in the analysis to 1399.

A generalized linear mixed effects analysis (Poisson distribution) was performed, with Type of Auditory Feedback, Type of Visual Feedback, and Disruption as independent variables. In order for the model to converge, the random effects structure was simplified; the interaction term was removed from the by-participant random slope and the random effect of item had an intercept only.

Model comparison revealed a significant interaction between auditory feedback, visual feedback, and disruption ($D = 17.54, p < .01$). Table 3.13 shows the fixed effects coefficients, standard errors, and z-values for the model with the significant interaction. The interaction is visualized in Figure 3.10, with the predicted
<table>
<thead>
<tr>
<th>Term</th>
<th>Coefficient</th>
<th>(Std. Error)</th>
<th>z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>−1.05</td>
<td>(0.24)</td>
<td>−4.40</td>
</tr>
<tr>
<td>auditory(DAF)</td>
<td>−0.01</td>
<td>(0.28)</td>
<td>−0.03</td>
</tr>
<tr>
<td>visual(picture)</td>
<td>−0.08</td>
<td>(0.30)</td>
<td>−0.25</td>
</tr>
<tr>
<td>visual(video)</td>
<td>−0.78</td>
<td>(0.33)</td>
<td>−2.41</td>
</tr>
<tr>
<td>disruption</td>
<td>0.18</td>
<td>(0.19)</td>
<td>0.95</td>
</tr>
<tr>
<td>auditory(DAF):visual(picture)</td>
<td>0.64</td>
<td>(0.39)</td>
<td>1.62</td>
</tr>
<tr>
<td>auditory(DAF):visual(video)</td>
<td>1.44</td>
<td>(0.41)</td>
<td>3.54</td>
</tr>
<tr>
<td>auditory(DAF):disruption</td>
<td>0.73</td>
<td>(0.22)</td>
<td>3.27</td>
</tr>
<tr>
<td>visual(picture):disruption</td>
<td>−0.00</td>
<td>(0.25)</td>
<td>−0.01</td>
</tr>
<tr>
<td>visual(video):disruption</td>
<td>0.44</td>
<td>(0.26)</td>
<td>1.67</td>
</tr>
<tr>
<td>auditory(DAF):visual(picture):disruption</td>
<td>−0.47</td>
<td>(0.32)</td>
<td>−1.48</td>
</tr>
<tr>
<td>auditory(DAF):visual(video):disruption</td>
<td>−1.01</td>
<td>(0.32)</td>
<td>−3.19</td>
</tr>
</tbody>
</table>

**Table 3.13:** Fixed effects for the model of speech errors with the significant interaction between auditory feedback, visual feedback, and DAF disruption. (NB: Since the model used a Poisson distribution, the coefficients are in expected log counts.)

As can be seen in Figure 3.10 there is considerable overlap of the confidence intervals, especially in the DAF conditions. The greatest distinction can be found at the low end of the disruption range in the NAF conditions. For sentences produced with NAF, the model predicts that speakers who are minimally disrupted by DAF will have a reduction in speech errors with dynamic visual feedback compared to static and no visual feedback.
Figure 3.10: Interaction between DAF disruption, auditory feedback, and visual feedback. Dotted lines represent 95% confidence intervals around predicted values. (NB: The predicted values are based on the fixed effects of the model only, and not the random effects.)
3.4.4 Discussion

The results of this experiment showed an effect of dynamic visual feedback on utterance duration; however, it was in the opposite direction to the initial prediction. In the DAF conditions, utterance duration increased with dynamic visual feedback compared to the conditions where visual feedback was static or absent. While the expected increase in speech errors in the DAF conditions compared to the NAF conditions was found, the initial analysis failed to find an effect of the different types of visual feedback. However, an additional analysis found that speech errors in the NAF conditions were reduced with dynamic visual feedback for those speakers who were minimally disrupted by DAF. Static visual feedback did not have an effect on either of the dependent variables.

The increase in utterance duration in the DAF-Video condition is consistent with the results from Experiment 1, and the interpretation of those results is also applicable here. It is possible that the duration of the closed portion of syllables was increased with DAF, in line with the articulatory findings from Zimmermann et al. (1988), and that the dynamic visual feedback reinforced this, resulting in a further increase to this portion of the syllable. While the differences were small, more prolongations were identified during error coding in the DAF conditions with visual feedback compared to no visual feedback; prolongations accounted for 22% of the total error in DAF-Video, 20% in DAF-Picture, and 17% in DAF-NoVideo. One possible caveat with this explanation is that the prolongations coded in Experiment 1 did not follow this pattern; prolongations accounted for more of the total errors in DAF-Picture condition (22%) than in the DAF-Video condition (18%) in Experiment 1. If the increased utterance duration is in part due to more and/or longer prolongations when DAF is produced with dynamic visual feedback, as suggested above, then we might have expected to see more prolongations in the Experiment 1 DAF-Video condition than in the DAF-Picture condition, which was the pattern observed in Experiment 2. However, it is important to keep in mind the coding criterion used for prolongations. Prolongations were defined as “abnormal and/or incongruous prolongation of a segment within a word” (Table 3.2). Thus, if an utterance was produced very slowly with many prolongations throughout, any one instance of prolongation would not have been “abnormal and/or incongruous,” and
as such would not have been coded as a prolongation. The coded prolongations represent extreme cases of prolongation within a given utterance. All this is to say, that even though there were fewer prolongations coded in the DAF-Video condition compared to the DAF-Picture condition for Experiment 1, there may still have been longer syllable closure durations which contributed to the overall increase in utterance duration for the DAF-Video condition. A quantitative method for identifying prolongation, or including a measure of segment durations as per Chesters et al. (2015), would clarify this issue.

The predicted reduction in speech errors with visual feedback was not observed in the main analysis. However, a further analysis which tested whether there was a differential effect of visual feedback based on the degree of DAF disruption a speaker experienced did reveal an effect (where ‘disruption’ was the mean number of errors a speaker made in the DAF-NoVideo condition). Speakers who produced minimal speech errors with DAF showed a reduction in speech errors when dynamic visual feedback was paired with NAF. Jones and Striemer (2007) also found that visual feedback reduced speech errors for participants classified as “low-disruption,” although this was observed when visual feedback was paired with DAF, not NAF. They interpreted this to mean that low-disruption speakers were those who could largely ignore their auditory feedback and instead focus on feedback that was synchronous with their actual output (i.e. somatosensory and visual feedback).

The different results in these two experiments could be due to differences in the experimental design and analysis. Jones and Striemer (2007) included a condition where participants repeated the auditorily presented sentences while also reading them on the monitor. In the analysis of the low-disruption group in Jones and Striemer (2007), it was the difference between this condition and the visual feedback condition which was significantly different (when paired with DAF). They failed to find a significant difference within this subset of speakers for the conditions which would be equivalent to DAF-NoVideo and DAF-Video in the present experiment. In this respect, the results for the two experiments could be considered similar, since the analyses in this chapter also failed to find a difference here. As for the different results reported for the NAF comparisons, this could be due to the different analyses used. Unlike the analysis presented in this chapter, which used a
continuous measure of disruption, Jones and Striemer used a median split to create two categories of speakers: low- and high-disruption (where ‘disruption’ was operationalized in the same way as the experiments in this chapter). A median split treats all values in a category as being equal, whether a value is close to the median or far away, thus there is some loss of information. As can be seen in Figure 3.10, the reduction in errors in the NAF-Video condition at the lowest end of the disruption range was quite modest. If a similarly small reduction occurred in Jones and Striemer’s data, this pattern could have been obscured once all participants below the median disruption level were grouped together.

Part of the interpretation proposed by Jones and Striemer (2007) for their low-disruption results is still relevant for the results of the present experiment; namely, speakers who are minimally disrupted by DAF most likely rely less on auditory feedback than speakers who experience extensive disruption. This interpretation is also supported by the results of the DAF experiment by Fabbro and Daró (1995) with simultaneous interpreters. According to Chon et al. (2013), these minimally disrupted speakers could be considered more verbally proficient, with more accurate or stable feedforward control of speech. The results of Experiment 2 also raise the possibility of an additional requirement for visual feedback to have an effect. When dynamic visual feedback is added speech output can become more fluent, but this may depend on all sources of feedback being synchronous (i.e. in the NAF condition); the results suggest that it is not sufficient for the visual feedback to be synchronous with just the somatosensory feedback but not the auditory feedback. While this interpretation is at odds with the results from Jones and Striemer, there is some support for this in the results from Chesters et al. (2015) showing that speech production was disrupted to a similar degree by a variety of visual feedback delays, regardless of whether there was synchrony with the auditory feedback (i.e. by also delaying the auditory feedback). In all cases there was some degree of asynchrony between the auditory, somatosensory, and visual feedback.

In considering such an interpretation of the results, it should be noted that the temporal requirements for audio and visual signal alignment have been investigated extensively in the context of speech perception. A visual influence on auditory perception does not require the two signals to be synchronous. For example, while the McGurk effect is strongest when the signals are temporally aligned (Munhall
et al., 1996) there is still evidence of a (weakened) visual influence when the visual signal leads the audio signal by as much as 360 ms (Jones and Jarick, 2006). The temporal window of visual influence is typically described as asymmetric; a visual signal that leads an audio signal is tolerated more than an audio signal that leads a visual signal, although the exact range of the window may depend on the type of stimuli used (Schwartz and Savariaux, 2014).

There may be a similar tolerance for some amount of temporal asynchrony between speech feedback signals, but perhaps not to the same extent as seen in perception. There is evidence to suggest that temporal misalignments among feedback signals are not well tolerated. The clearest example of this is the disruptive effects of DAF. Additionally, the targeted delays described in the study by Cai et al. (2011) resulted in the production of longer durations that extended beyond the point of the temporal perturbation. And shadowing tasks that involve asynchronous audiovisual stimuli do not result in the accuracy improvements seen with synchronous audiovisual stimuli (Reisberg et al., 1987). Given this, it is possible that a greater degree of temporal cohesion among multimodal signals is more important in the context of speech feedback than speech perception. This issue is returned to in Chapter 5.

The analysis of Experiment 2 suggests that static visual feedback does not elicit changes in speech production. One potential concern with the static visual feedback presented in this experiment is the fact that it was an image of a face in a non-speech posture. In the Calvert and Campbell (2003) experiment described in Section 3.4, the static visual images were from the peak of the syllable and the baseline image was a face with a neutral expression, with the target letter superimposed on the lips. For the comparison of the baseline to the static images conditions, it was noted that the static images resulted in greater activation of the face fusiform region (which is important for facial recognition), despite the fact that both images were of a stilled face. This baseline condition is similar to the static visual feedback used in this experiment, and so raises the question of the extent to which this feedback was processed as being relevant for a speech task.

Assuming that the static visual feedback was treated by speakers as being relevant for speech, there are still other possible reasons why this form of visual feedback did not affect speech output in the manner predicted (i.e. by minimizing dis-
fluencies, but to a lesser degree than dynamic visual feedback). One difficulty in presenting static visual feedback during the production of whole utterances, is that it is not possible to present static feedback which matches the auditory signal (e.g. producing [u] and seeing a still image of the face articulating [u]). The neutral expression with mouth closed that participants were presented may have had varying effects depending on what was being articulated. For example, this static form may have facilitated production during closed or labial sounds, reducing speech errors, but hindered production during open sounds, increasing speech errors. Alternatively, Campbell (1996) suggests that a static face paired with an auditory speech signal could be parsed as two separate events rather than a single event, thus diminishing the effects of the visual signal on speech processing. The results of this experiment suggest that the production of longer utterances is more sensitive to dynamic visual information; this contrasts with the examples described in Section 3.4 which involved static visual information influencing single syllables or short multisyllabic words.

The results presented in this chapter suggest that dynamic, but not static, visual feedback influences speech production. This effect was not consistent across the conditions and measures, however. Speech produced with DAF exhibited a further increase in utterance duration when paired with dynamic visual feedback. In contrast, dynamic visual feedback had a facilitatory effect in terms of speech errors (i.e. speech errors were reduced), but only when paired with NAF and only for those speakers who were minimally disrupted by DAF. One possible interpretation of this is the importance of temporal cohesion among feedback signals.

### 3.5 General discussion

The experiments presented in this chapter tested whether visual feedback can enhance speech production output during a DAF task. Experiment 1 randomized the presentation of static and dynamic visual feedback within each block in order to make it less likely that participants would respond strategically, which is common in response to DAF. Utterance duration unexpectedly increased when dynamic visual feedback was paired with both normal and delayed auditory feedback and there were no changes in the number of speech errors between the two visual feed-
back conditions. The failure to find a facilitative effect of dynamic visual feedback may have been due to the random presentation of the types of feedback.

Experiment 2 grouped the different types of visual feedback into consistent blocks, to contrast with Experiment 1. Additionally, a condition with no visual feedback was added in order to more carefully compare the effects of static and dynamic visual feedback. Once again, utterance duration increased with dynamic visual feedback, but in this experiment it was only when paired with DAF. While the main analysis did not find an effect of visual feedback on speech errors, an additional analysis which included the degree of DAF disruption experienced by the participant as a predictor showed that speech errors were reduced for participants who were least disrupted by DAF when dynamic visual feedback was paired with normal auditory feedback. This was interpreted to mean that verbal proficiency is relevant for the use of visual feedback during the production of whole utterances, and also that temporal cohesion among signals may have more stringent requirements in feedback than it does in perception. The results also suggest that it is the time-varying component of visual feedback, rather than the static form, which is important in enhancing speech output. A more detailed comparison of the two experiments is presented below.

The increased duration observed when dynamic visual feedback was paired with NAF in Experiment 1 was attributed to a possible effect of segmental duration increases, following results from Chesters et al. (2015). However, this may also have been due to the greater number of pauses which occurred in Experiment 1 compared to Experiment 2, as shown in Figure 3.4 compared to Figure 3.8. The number of pauses identified during error coding for all conditions in both experiments is shown in Table 3.14. Since there were an additional four stimuli in Experiment 1, the pauses counts are represented as pauses per 100 sentences to facilitate comparison between the experiments. In Experiment 1 there were more than twice as many pauses in the dynamic visual feedback conditions, for both NAF and DAF, compared to Experiment 2. For the DAF-Picture condition there slightly less than twice as many pauses in Experiment 1 than Experiment 2, and for the NAF-Picture condition there were a similar number of pauses in the two experiments. The greater number of pauses in Experiment 1 may have contributed to the longer utterance durations when dynamic visual feedback was presented.
Table 3.14: Number of pauses per 100 sentence for each condition in Experiment 1 and Experiment 2.

<table>
<thead>
<tr>
<th>Condition</th>
<th>No Video</th>
<th>Picture</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>NAF –</td>
<td>6.1</td>
<td>8.9</td>
</tr>
<tr>
<td></td>
<td>DAF –</td>
<td>26.7</td>
<td>32.5</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>NAF</td>
<td>7.0</td>
<td>5.7</td>
</tr>
<tr>
<td></td>
<td>DAF</td>
<td>15.2</td>
<td>15.4</td>
</tr>
</tbody>
</table>

One point to note on this issue is that all pauses, which were quite long in some instances (up to 2000 ms in DAF conditions), were included in the measure of utterance duration. Duration was calculated with pauses included in order to be consistent with previous studies looking at visual feedback during a DAF task. While the previous studies using this paradigm don’t explicitly state that there were long pauses in the data, the reported mean durations of utterances and their constituent segments in Chesters et al. (2015) suggest that pauses on the order of 750 ms (on average) occurred in that experiment. While pauses are undoubtedly a common feature of speech produced with DAF, their inclusion in measures of utterance duration may obscure some of the more subtle effects of visual feedback. A measure such as articulation rate, which calculates syllables per second minus pauses (e.g. Chon et al. 2013, Stuart et al. 2002), may be more insightful for measuring differences between visual feedback conditions, and is recommended for future research.

This increase in the number of pauses in Experiment 1 compared to Experiment 2 was echoed in the main dependent variables. Overall, Experiment 1 resulted in greater disfluency, with both the utterance duration means and speech error means having higher values than in Experiment 2. The means and standard errors are reproduced in Table 3.15 for ease of comparison. The biggest differences are seen for the DAF conditions. The mean utterance duration was greater by 395 ms and 345  

3These studies also don’t report whether pauses of a certain duration were excluded from the analysis.
Experiment 1

<table>
<thead>
<tr>
<th>Utterance duration (ms)</th>
<th>NAF</th>
<th>–</th>
<th>2235 (66)</th>
<th>2253 (66)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAF</td>
<td>3479 (76)</td>
<td>3539 (66)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. speech errors per utterance</td>
<td>NAF</td>
<td>–</td>
<td>0.36 (0.06)</td>
<td>0.39 (0.06)</td>
</tr>
<tr>
<td>DAF</td>
<td>1.38 (0.07)</td>
<td>1.41 (0.08)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Experiment 2

<table>
<thead>
<tr>
<th>Utterance duration (ms)</th>
<th>NAF</th>
<th>2188 (62)</th>
<th>2125 (56)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAF</td>
<td>3084 (90)</td>
<td>3194 (78)</td>
<td></td>
</tr>
<tr>
<td>No. speech errors per utterance</td>
<td>NAF</td>
<td>0.39 (0.06)</td>
<td>0.29 (0.05)</td>
</tr>
<tr>
<td>DAF</td>
<td>0.98 (0.07)</td>
<td>0.97 (0.07)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.15: Means and standard errors of the means for the dependent variables in Experiment 1 and Experiment 2.

ms for the DAF-Picture and DAF-Video conditions, respectively, of Experiment 1 compared to Experiment 2. The mean number of speech errors per utterance was greater by 0.38 for the DAF-Picture condition, and 0.44 for the DAF-Video condition of Experiment 1 compared to Experiment 2. Longer durations were also observed for the NAF conditions of Experiment 1 compared to Experiment 2, but these differences were smaller than the differences between the DAF conditions of the two experiments. One exception to this pattern is the NAF-Picture condition; there were fewer speech errors on average in Experiment 1 compared to Experiment 2.

One interpretation of this overall greater disfluency is that Experiment 1 was more difficult than Experiment 2, possibly due to the randomized presentation of the different types of visual feedback. If this is the case, the failure to find a facilitative effect of dynamic visual feedback in Experiment 1 may have been due to task difficulty. In perception, audiovisual integration has been shown to be somewhat
sensitive to the demands of increased cognitive load. Using a McGurk paradigm, the addition of non-speech perceptual monitoring task (in either the auditory or visual domains) has been shown to reduce the influence of visual information (Alsius et al., 2014, 2005). A more modest reduction in the McGurk effect has also been reported when cognitive load is increased by adding a working memory task (memorizing a string of numbers) (Buchan and Munhall, 2012). There are similarities between these tasks, which were designed to increase difficulty, and conditions of Experiment 1, which involved: memorization of the stimuli to be repeated, additional auditory feedback monitoring demands introduced by the alternating blocks of delayed and non-delayed auditory feedback, and monitoring of randomly presented static and dynamic visual feedback (an unusual source of speech feedback). The conditions of Experiment 2 involved the first two of these, but each block presented just one type of visual feedback. This may have reduced the difficulty of the task enough such that visual feedback could be observed. However, Experiment 2 is still a challenging task, and this may be part of the reason why the reduction in speech errors with dynamic visual feedback was primarily found for more verbally proficient participants.

Finally, with regard to the looking variable: As discussed in Section 2.4, each item was coded for whether the participant was looking at the monitor during the repetition of the stimulus, after it was noted that many participants looked away from the monitor for a large number of items. This information was included in the analyses as a control variable. Similar procedures—or alternatives such as the use of catch trials—are not reported in previous studies (Chesters et al., 2015; Jones and Striemer, 2007; Tye-Murray, 1986). One concern with this is that the analyses of speech errors in Experiment 1 and 2, and the analysis of utterance duration in Experiment 1, showed that the looking variable was a significant factor in these models; participants produced more errors and greater durations when they were looking away from the monitor compared to when they were looking at the monitor. It is possible that some of the patterns reported in earlier work could have been influenced by trials in which participants were not looking at the visual feedback, which would contribute to some of the variability across the studies. It also makes it difficult to interpret reported effects of visual feedback if there’s a chance that participants were not actually looking at the visual feedback.
Why might participants have looked away, or closed their eyes, in such high numbers during this task? One possible reason is that participants felt uncomfortable watching themselves; at the end of the experiments, many participants commented on how strange or funny it was to see themselves talk. However, this reason is unlikely, given that participants consistently looked at the monitor during the bite block task with visual feedback, which is presented in the next chapter. Perhaps the most likely reason is the challenging nature of the task, as discussed above. This is supported by the fact that there were more sentences in which participants looked away from the monitor during Experiment 1, the more challenging of the experiments, than Experiment 2. Additionally, production was more disfluent when participants were not looking at the monitor, suggesting that they looked away when they felt particularly challenged by the task. This interpretation is consistent with work showing that gaze aversion (i.e. looking away from an interlocutor) increases in response to increased difficulty of a cognitive task, and this happens in face-to-face communication as well as video-conferencing (Doherty-Sneddon and Phelps, 2005).

The two experiments in this chapter explored the effect of visual feedback on a delayed auditory feedback (DAF) task. The results of Experiment 1 suggest that visual feedback does not enhance speech production in a DAF task which varies the presentation of visual feedback. However, when the different types of visual feedback (dynamic, static, no visual feedback) were presented in consistent blocks as in Experiment 2, there was a modest reduction in speech errors. These improvements occurred with dynamic, but not static, visual feedback, suggesting that it is the time-varying properties of visual feedback which have the potential to elicit changes in speech production. This is consistent with the hypothesis that it is the temporal properties of visual feedback that establish compatibility with other sources of speech feedback. The finding that enhanced fluency with dynamic visual feedback (found only for the most verbally proficient speakers) occurred during productions with normal, rather than delayed, auditory feedback, suggests that temporal compatibility between feedback signals may include a requirement for synchrony. That is, feedback can enhance productions only when all signals are aligned; it may not be sufficient for only a subset of the feedback signals to be aligned (e.g. only somatosensory and visual feedback in the case of delayed
auditory feedback). This issue is returned to in Chapter 5.

The experiment presented in the next chapter (Chapter 4) was designed with these issues in mind. The next experiment is a bite block task, in which participants repeat single words with and without a bite block, and with and without visual feedback. An oral perturbation was used as this difficult speaking task makes it more likely that speakers will rely on sensory feedback (Lane et al., 2005) but it does not interfere with the temporal synchrony of the feedback signals. During the experiment participants are presented with consistent blocks of visual feedback, as in Experiment 2 from the present chapter. Finally, to reduce the difficulty of the task, participants are required to repeat single words rather than whole sentences.
Chapter 4

Visual feedback during a bite block task

4.1 Overview

This chapter reports on an experiment that examined the effect visual feedback has on the production of vowels. Participants produced monosyllabic words in four conditions that varied in terms of the presence or absence of visual feedback and an oral perturbation. The acoustic analysis provides evidence that visual feedback can enhance the distinction between vowel categories and can minimize the variability within vowel categories. Optical flow analysis of video recordings revealed considerable inter-speaker variation in lower face motion; however, a subset of participants did produce greater magnitudes of motion during normal production with visual feedback. While there was a modest positive correlation between facial motion and acoustic contrast overall, the participants who produced vowels with greater magnitudes of motion in the presence of visual feedback tended to exhibit less acoustic contrast than participants who produced vowels with smaller magnitudes of motion. Overall, these results support the hypothesis that visual feedback can enhance speech output, but they also raise questions concerning the relationship between articulation and acoustics in achieving this enhancement.
4.2 Introduction

4.2.1 The effect of visual information on phonological contrasts

Visual information has been shown to be relevant to the acquisition of phonological contrasts in two areas: altered speech production in the visually impaired suggests that visual information is involved in normal native language acquisition and non-native speech production that is guided by visual speech information suggests that visual information can be used to learn second language phonological contrasts.

Ménard and colleagues (2015; 2009; 2013) have documented the differences between congenitally blind and sighted French speakers in the production of vowels. Participants were required to produce each of the ten French oral vowels in the carrier phrase \textit{V comme pVpa} (\textit{V as in pVpa}), with acoustic and articulatory measures made of the initial, sustained vowel. Euclidean distances between specific vowel pairs were measured to determine acoustic vowel contrast, with the mean of these distances representing the global between-category contrast for the vowel space. Blind speakers were found to produce less between-category acoustic contrast than sighted speakers (Ménard et al., 2009). When this measure was used to compare specific phonological feature contrasts rather than the vowel space as a whole, the same decrease in acoustic contrast was observed for blind speakers, especially for contrasts that differed in terms of both rounding and place of articulation (e.g. /i/-/u/), rather than just one of these dimensions (e.g. /i/-/y/) (Ménard et al., 2013). The amount of dispersion within each vowel category, which “reflects the precision with which a specific goal is reached” (2013, p. 2984), was also assessed, by measuring the Euclidean distance from each token (in F1xF2xF3 space) to the mean of its category. Blind speakers’ vowel categories had greater dispersion than sighted speakers’ vowel categories (Ménard et al., 2013). Additionally, tongue position and curvature from ultrasound images and upper lip protrusion from video images were measured. Blind speakers produced smaller lip protrusion distances and greater tongue backing and tongue curvature compared to sighted speakers (Ménard et al., 2013).

The results from these studies suggest that visual speech information plays a role in the implementation of phonological contrasts, affecting both the degree of
contrast achieved and the manner in which the contrast is implemented. When visual information is not available during development, acoustic vowel contrast is reduced and the individual tokens of a vowel category are less tightly clustered. Ménard et al. (2013) suggest that this may be due to greater articulatory variability; they make a comparison to sighted speakers, suggesting that access to visual information during perception would help to reduce variability by providing additional information that could be used to guide their articulatory targets. These studies also show that for blind speakers the extent of movement of visible articulations is reduced, while non-visible articulations are enhanced.

In addition to this example of the absence of visual speech information impacting vowel contrasts, visual speech information has been investigated in the context of second language (L2) acquisition. As expected, given the enhanced perceptual accuracy for audiovisual speech reported in other difficult listening conditions (e.g. speech in noise [Sumby and Pollack, 1954]), audiovisual speech, compared to audio-only, also improves the perception of L2 vowel contrasts. For example, in a speeded syllable classification task which varied in terms of presence or absence of the Catalan vowel contrast /e/-/e/, only Catalan-dominant bilinguals showed sensitivity to the contrast when the stimuli were presented auditorily, but both Catalan-dominant and Spanish-dominant bilinguals were sensitive to the vowel contrast during audiovisual presentation of the stimuli (Navarra and Soto-Faraco, 2007). Improvements in producing L2 contrasts are also observed when visual speech information is made available. Computer-assisted pronunciation systems have been used to compare the effects of presenting different types of visual speech information. Using native speaker ratings of production accuracy as a measure, Massaro et al. (2008) reported improvements in the production of the Mandarin contrast /i/-/y/ when participants completed training while seeing a front view of lip motion produced by a talking head. Similar improvements were not found when participants learned the Arabic contrast /k/-/q/ with a sagittal view of the talking head which also displayed the internal articulators. Given that this result could have been due to the different phoneme contrasts rather than the type of visual feedback, there is scope for further exploration. Katz and Mehta (2015) took up this question in the context of real-time visual feedback, presenting English-speaking participants with a visualization of their moving tongue (derived from EMA sensors) within a talking
head while producing a voiced palatal stop. While articulatory accuracy—measured as the number of successful attempts at “hitting” an articulatory target displayed in the oral cavity of the talking head with a marker displayed on the tongue—improved after training, the analysis of the spectral burst in the acoustic recordings revealed considerable inter-speaker variability.

This research suggests that visual speech information plays a role in shaping vocalic and consonantal contrasts in the context of both development and the acquisition of non-native contrasts by adults. The experiment presented in this chapter tests whether providing visual feedback can also change adult speakers’ native vowel contrasts.

4.2.2 The effect of oral perturbations on vowel production

Bite block perturbations have been used in a range of research to investigate multimodal sensory processing. Early work showed that compensation for a bite block is almost immediate, and for the English vowels measured the acoustic compensation is nearly complete (Fowler and Turvey, 1980), suggesting a limited role for auditory feedback in the restructuring of the necessary articulatory dynamics. However, later work showed that immediate compensation may be much less, especially for consonants, but there is improvement over time (McFarland and Baum, 1995). Lane et al. (2005) aimed to adjudicate this issue by looking at the interaction between hearing status and oral perturbation. Participants included adult cochlear implant users (age at implantation ranged from 28 to 78) tested before and after implantation. The experiment tested the effect of the bite block with both masked and normal auditory feedback, and acoustic vowel dispersion (the “spread” of the tokens within a given category) and vowel contrast (the degree of separation between vowel categories) were measured. The results suggest that compensation for a bite block does involve use of auditory feedback. The bite block increased vowel dispersion, but this increase was greatest prior to implantation than it was when re-tested with the bite block one year after implantation, suggesting a role for long-term experience with auditory feedback in compensating for an oral perturbation.

In terms of articulatory reactions to oral perturbations, high inter-speaker vari-
ability is often reported, with compensation ranging from complete to absent (Ménard et al., 2008; Savariaux et al., 1995). In a comparison of blind and sighted speakers’ compensation for a lip-tube perturbation during the productions of French /u/, Ménard et al. (2015) found that blind speakers produced a greater degree of tongue curvature than sighted speakers in response to a lip-tube perturbation during the production of French /u/ (Ménard et al., 2015). This finding is consistent with the study described in the previous section, showing that blind speakers tend to produce more extreme non-visible articulations than sighted speakers (Ménard et al., 2013).

In the experiment presented in this chapter, a bite block paradigm was used in combination with the presentation of visual feedback during speech production. Following Lane et al. (2005) and Ménard et al. (2013), the acoustic contrast and dispersion of vowels were analyzed. These measures are defined in more detail in Section 4.3.4.

### 4.2.3 Hypothesis and predictions

The general hypothesis of this dissertation is that visual feedback, which has time-varying properties that make it compatible with other sources of speech feedback (auditory and somatosensory), will enhance speech output. The visual speech signal is a rich source of information, with jaw motion (Johnson et al., 1993) and lip position (e.g. Fromkin, 1964; Montgomery and Jackson, 1983) being particularly relevant for vowels. When this modality is not available during development, because of congenital blindness, the acoustic and articulatory properties of vowels are changed relative to speakers who have access to visual speech information (Ménard et al., 2009, 2013). Given this previous research, one way in which speech output is predicted to be enhanced is in terms of the production of vowel contrasts. Perturbing speech production with a bite block increases reliance on feedback (Lane et al., 2005), thus visual feedback is predicted to enhance speech output to a greater extent during bite block speech compared to normal (non-perturbed) speech.

Consistent with Lane et al. (2005) and Ménard et al. (2013), the following predictions were made. (A graphical summary of the predictions is presented in Figure 4.1.) Vowel contrast is predicted to be reduced during bite block production.
and increased with visual feedback. The visual feedback effect is predicted to be greater during bite block production compared to normal speech production. Vowel dispersion, which is used here to refer to the spread of tokens within a vowel category rather than the distance between vowel categories as in Dispersion Theory (Flemming, 1995), is predicted to be increased during bite block production and reduced with visual feedback. The visual feedback effect is predicted to be greater during bite block production compared to normal speech production. In line with the articulatory results from Ménard et al. (2013) showing articulatory differences between blind and sighted speakers during vowel production, visual feedback was predicted to result in comparable changes. The analysis of this experiment focused on visible articulatory differences since this aligns with what participants were able to see during the presentation of visual feedback. A global measure of lower face magnitude of motion was used, with the magnitude predicted to be reduced during bite block production and increased with visual feedback. Since the bite block will limit jaw opening, the visual feedback effect may be observed more during production without the bite block.
Figure 4.2: The canonical positions in the vowel space for the General American English monophthongs used in the experiment (adapted from Ladefoged (2006)). Each vowel was produced in a monosyllabic word of the form /hVd/ (represented orthographically next to the appropriate vowel.)

4.3 Methods

4.3.1 Participants

Thirteen students (mean age: 24.2 years (sd: 10.6); 5 males) recruited from The University of New Mexico’s Linguistics 101 course participated in the experiment. Participants were compensated for their time with extra credit toward their final grade. All participants self-reported having normal hearing and normal or corrected-to-normal vision, and all were native English speakers.

4.3.2 Stimuli

Seven monosyllabic words—of the form /hVd/—served as stimuli for participants’ production, and the vowel in each word was one of seven American English monophthongs. Figure 4.2 shows the vowel positions in the vowel space.
4.3.3 Procedure

The experiment, which required participants to repeat words in four different feedback conditions, followed the procedure outlined in Section 2.4. A within-participants design with two independent variables was used. These variables manipulated the type of oral perturbation and visual feedback participants received during speech production, with the variables being either absent or present. When the oral perturbation was absent, participants produced speech as they would normally. For the oral perturbation, participants produced the stimuli while holding a bite block between their upper and lower right molars. The bite block was a disposable pair of wooden chopsticks, approximately 8 mm in diameter. An example of this configuration is shown in Figure 4.3. Visual feedback consisted of a real-time video of the utterance being produced by the participant. When visual feedback was absent, participants looked at a fixation point on the monitor. Each participant produced repetitions of the stimuli in the four conditions outlined in Table 4.1.

The stimuli were presented multiple times in each condition: each condition contained four blocks, each of which contained two repetitions of each stimulus. Stimulus order was randomized in each block. Condition order was pseudo-randomized across participants; participants always completed the conditions without the bite block before the conditions with the bite block, but the order of the visual feedback conditions was randomized within each of these subgroups. The oral perturbation conditions were ordered in this way to avoid any possible compensation or adaptation effects which could occur when shifting from the bite block.
Figure 4.3: Bite block configuration used in the oral perturbation conditions. Participants held the chopsticks between their upper and lower right molars while producing the stimuli. During speech production participants looked straight ahead at the visual feedback on the monitor with the chopsticks typically lying parallel to the microphone. (Note that this is not the case in this image. This video still was chosen to more clearly show the bite block; it came from a period between blocks while the participant turned to the side.)

Conditions to the no bite block conditions. This resulted in the four condition orders presented in Table 4.2.

In total each participant produced 14 words per block, and thus 56 words per conditions and 224 words for the experiment, yielding a total of 2912 words across all 13 participants. 2 words were discarded due to technical errors during the recording of the experiment, leaving 2910 words in the analysis.

1In addition to these stimuli, a set of sibilant stimuli (seat, sheet, sort, short) were also included. In total then, each participant produced 22 words per block, and thus 88 words per condition and 352 words for the experiment.

2In a pilot study for this experiment, participants produced twice as many repetitions of each word. However, this proved to be too taxing for the participants during the bite block conditions. To avoid participants becoming too fatigued during the experiment, the number of repetitions per condition was reduced to two.
Table 4.2: The four condition orders used in the experiment.

<table>
<thead>
<tr>
<th>Order 1</th>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Condition 3</th>
<th>Condition 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoBB-NoVideo</td>
<td>NoBB-Video</td>
<td>BB-NoVideo</td>
<td>BB-Video</td>
<td></td>
</tr>
<tr>
<td>NoBB-Video</td>
<td>NoBB-NoVideo</td>
<td>BB-NoVideo</td>
<td>BB-Video</td>
<td></td>
</tr>
<tr>
<td>NoBB-NoVideo</td>
<td>NoBB-Video</td>
<td>BB-Video</td>
<td>BB-NoVideo</td>
<td></td>
</tr>
<tr>
<td>NoBB-Video</td>
<td>NoBB-NoVideo</td>
<td>BB-NoVideo</td>
<td>BB-NoVideo</td>
<td></td>
</tr>
</tbody>
</table>

4.3.4 Measures

4.3.4.1 Acoustic measures

Two acoustic measures were used: vowel contrast and vowel dispersion. Vowel contrast refers to the distance between vowel categories, and aims to capture the spread of the vowel space. Vowel dispersion refers to the distance between individual vowels and the mean of their category, and aims to capture the spread of vowel categories.

The audio recordings were segmented, annotated, and measured using Praat (Boersma and Weenink, 2014) and FAVE (Rosenfelder et al., 2011). FAVE (Forced Alignment and Vowel Extraction) includes two programs which enable automatic alignment of orthographically transcribed data and acoustic recordings, and automatic extraction of vowel formant measures. The command-line versions of these programs were used for the analysis of the present data set. A Praat script was used to create a TextGrid with word boundaries and phoneme labels. FAVE-align was then used to automatically add phoneme boundaries for each word. These boundaries were checked and manually corrected if required. During the manual check and correction phase, vocalic onsets and offsets were identified by the following criteria: 1) Vocalic onsets were identified by the onset of periodicity in the waveform and the onset of harmonic structure in the spectrogram. In cases where these did not align due to temporal smearing in the spectrogram, the boundary was placed at the onset of periodicity. 2) Vocalic offsets were identified by an abrupt
intensity decrease in the waveform and the offset of high frequency components in
the spectrogram. In some cases the offset could also be identified by the offset of
periodicity.

Formant analysis was carried out using FAVE-extract as an interface for Praat. Before running FAVE-extract, formant settings were chosen for each participant. To do this, the spectrograms from a random word in each block were viewed in tandem with Praat’s formant tracker, and the settings were adjusted to optimize the formant tracker’s performance.\textsuperscript{3} The two settings adjusted were the number of formants and the upper limit of the formant search range. Once these values were determined, the FAVE-extract settings were specified: Praat’s formant prediction algorithm along with the pre-determined settings for each participant. Praat’s algorithm was chosen as the formant prediction method over the Mahalanobis prediction method as this enabled subsequent manual checking of the automatically extracted formant values in Praat. For each participant, formant values were extracted at three points during the vowel production—25%, 50%, and 75%—in order to capture some of the temporal dynamics.

Following Lane et al. (2005), the effects of the oral perturbation and visual feedback manipulations were assessed in terms of vowel dispersion and vowel contrast. Vowel dispersion refers to the spread of a vowel category, operationalized as mean Euclidean distances ($d$), and vowel contrast refers to the spread of the vowel space, operationalized as mean Mahalanobis distances ($D_M$).

Euclidean distance is the distance between two points in Euclidean space; in this case, the distance between a point and the average of all the points of that category. For example, the distance between each instance of [i] and the mean of all the tokens of /i/, measured in F1xF2 space. The Euclidean distance formula is given in Equation 4.1. This calculation was applied per participant per condition at the three time points during the vowel production. That is, for each participant, and at each of the three time points, the mean of each vowel category was calculated for each condition, and the distances between this mean and each token from that category were calculated.

\textsuperscript{3}The formant tracker had the most difficulty with /u/, so each token of who’d was also checked.
\[ d(V_i, V_j) = \sqrt{(F1_i - F1_j)^2 + (F2_i - F2_j)^2} \]

where \( V \) = a vowel category

\( i \) = the \( i \)-th vowel category

\( \bar{i} \) = the mean of the \( i \)-th vowel category

\( F1 \) = the first formant

\( F2 \) = the second formant

Mahalanobis distance is similar to Euclidean distance, however the formula takes into account the variance and covariance of the category a particular token is being compared to, not just the mean of that category (Lane et al., 2005). This measure has been used in a variety of speech studies ranging from automatic speech recognition, for mapping training data from one language to a new language (e.g. Sooful and Botha, 2001), to second language learning, for calculating the distance between L2 productions and the native language target (e.g. Kartushina et al., 2015). Multiple Mahalanobis distances were calculated for each token, as each token was compared to the distribution of each of the seven vocalic categories at three points in the vowels. Thus for each token, twenty-one distances were calculated (e.g. [i] \( \rightarrow \) /i/, [i] \( \rightarrow \) /ɪ/, [i] \( \rightarrow \) /æ/, [i] \( \rightarrow \) /æ/, [i] \( \rightarrow \) /ʌ/, [i] \( \rightarrow \) /ʌ/, at 25%, 50%, and 75% points in the vowel). The Mahalanobis distance formula is given in Equation 4.2.

\[ D_M = \sqrt{(x_i - \mu_j)^T S^{-1} (x_i - \mu_j)} \]

where \( x \) = a vector of F1 and F2 values

\( \mu \) = a vector of mean F1 and F2 values

\( i \) = the \( i \)-th vowel category

\( j \) = the \( j \)-th vowel category

\( (.)^T \) = the transpose operation

\( S^{-1} \) = the inverse covariance matrix

These distances were used to calculate the average vowel space (AVS) distance, using a similar procedure to Lane et al. (2005). AVS was the mean of the
Mahalanobis distances between vowel pairs, calculated as per for the following procedure:

1. For each token, the Mahalanobis distance to the distribution of each vowel category was calculated. This resulted in seven distances for each token. A visualization of this step is shown in Figure 4.4.

2. The AVS distance for each token was calculated by averaging over the seven distances from step 1. These token-specific AVS distances were used in the statistical models.

3. This preceding two steps were repeated at each measurement point in the vowel (25%, 50%, 75%)

There are two differences between this procedure and the one used by Lane et al. (2005). The Mahalanobis distance between a token and its own category (e.g. the distance between [i] and /i/) was measured and included in the calculation of AVS for the present analysis, but not in Lane et al.’s analysis. This distance was included in the calculation in order to capture all of the vowel category distributions in the AVS measure. Lane et al. calculated pairwise Mahalanobis distances, and then averaged over these distances to calculate AVS. This approach results in a reduced data set for the statistical analysis, as the AVS measure is calculated for a set containing all the different vowel pairs; that is, instead of there being an AVS measure for each token of a vowel, there is one AVS measure for a set of vowels. This can be problematic for within-participant analyses and for repeated measures (Karlsson and van Doorn, 2012). The approach described above is an attempt to avoid this reduction in the number of data points for the statistical analysis.

4.3.4.2 Articulatory measures

A measure of visible articulation, focusing on the lower face, was used for the analysis. Motion of the lower face was measured using optical flow analysis (OFA),

---

4The description of this calculation in Lane et al. (2005) is initially described as involving pairwise distances, although at one point in the paper what is described is similar to the procedure used in the present analysis: “Each repetition of a particular vowel on the ith trial is given a Mahalanobis distance to the distribution of each of the other vowels. The square roots of the distances were averaged for each group and listening condition” (2005, p. 1640).
a method for extracting 2D movement information from videos without the need for pre-defined measurement locations. The basic operation of the algorithm is described by (Barbosa et al., 2008b, p. 6-7):

Moving images are recorded as changes in the intensity (and color) values for the pixels in the image array that are influenced by the motion. The optical flow algorithm does not merely register the change of intensity from one image to the next for each pixel; rather it attempts to keep track of specific intensity values, corresponding to image objects as they change location within the pixel array. Thus, the algorithm assigns a motion vector consisting of a magnitude and a direction to each pixel based on where the intensity associated with that pixel in one image is located in the next image in sequence. The direction is simply the line from the first pixel to the second and the magnitude corresponds to the Euclidean distance between them. The array of motion vectors comprises the optical flow field.

Since the difference between adjacent frames involves a temporal difference, this displacement of intensity values can be represented as pixel velocity and thus the
inferred movement is in terms of velocity of motion.

The algorithm used in the present analysis is from Horn and Schunck (1981), as implemented by Barbosa et al. (2008b) in their FlowAnalyzer software. A single, continuous video recording of the experiment was made for each participant. Using FlowAnalyzer, a region of interest (ROI) was specified for each video recording. The ROI was positioned to capture the speaker’s lower face movements—particularly the mouth—throughout the experiment. As such, the top edge of the rectangular ROI was positioned near the tip of the speaker’s nose and the bottom edge was positioned at the top of the microphone (which was usually aligned with the bottom of the speaker’s chin). The outer edges of the ROI were positioned to include the maximal extent of mouth position along the transverse plane throughout the experiment. An example ROI is shown in Figure 4.5.

In determining the ideal ROI size and position, a number of issues were taken into consideration. The main concern was whether the ROI should encompass the whole lower face or just the mouth. The difficulty with using a mouth-only ROI is

5https://www.cefala.org/FlowAnalyzer/
that multiple ROIs must be created throughout the experiment, since the position of the speaker’s head, and thus mouth, moves between blocks and conditions. This means that an additional normalization procedure would need to be introduced in order to pool data across the multiple ROIs for each speaker.\(^6\)

In order to see the differences between the motion measured within the lower face ROI and the mouth ROI, an informal comparison of the two was made. OFA was performed on the video recording of the first block from each of the two NoBB conditions for one participant. As can be seen in Figure 4.6, the pattern across the vowels was essentially the same for the two ROIs, but with larger magnitudes for the mouth only ROI. Given that similar patterns of motion were extracted from the two test ROIs, the choice was made to use the lower face ROI for ease of data extraction. Large ROIs have also been successfully used in other contexts. When OFA is used to calculate correlations between multimodal speech signals, large ROIs (encompassing the whole lower face, for example) have been shown to perform on par with more targeted analyses using marker dots for capturing facial motion (Barbosa et al., 2008a). When used to measure tongue movements from ultrasound, similar results were obtained regardless of whether a wide or narrow ROI was used (Hall et al., 2015).

The output of OFA includes five vectors quantifying values and magnitudes in the horizontal and vertical directions, as well as a summed magnitude of motion (velocity) within the ROI. This summed magnitude is a single scalar value; it is the sum of the Euclidean magnitudes for each pixel in the ROI. These values are calculated for each frame-step in the video sequence. Following previous research (e.g. Barbosa et al., 2008a; Fuhrman, 2014; Hall et al., 2015), the summed magnitude of motion (MM) was used in the present analysis.

Since MM values are highly correlated between adjacent frame-steps, an averaging procedure was used to derive a single MM value for each token (Hall et al., 2015). Each mean MM was calculated from the frame-steps that occurred during the onset of the word (/h/) plus the first half of the vowel. This portion of the word

\(^6\)An alternative view is that it may not be necessary to normalize when using a mouth-only ROI. Creating multiple ROIs which only capture mouth movement is akin to using OFA as an object-tracker, and since the object changes shape the ROI will need to change accordingly (Adriano Vilela Barbosa, personal communication).
Figure 4.6: Comparison of mouth and lower face magnitudes of motion from two ROIs. Calculations were made from the first block of the NoBB-NoVideo condition (left) and the first block of the NoBB-Video condition (right) for one participant. In each of these blocks the participant produced two repetitions of each stimulus, so each point represents the mean of two repetitions.

was chosen in order to capture the maximum extent of the opening gesture of the target vowel. A Python script\textsuperscript{7} time-aligned the Praat TextGrid boundaries with the OFA frame-step sequences, then extracted the OFA vectors and calculated the relevant means.

4.4 Results

The F1xF2 vowel spaces, as measured at the 25%, 50%, and 75% points in the vowel, are shown in Figures 4.7, 4.8, and 4.9, respectively. The statistical results for the acoustic and articulatory measures are discussed below.

\textsuperscript{7}The original Python script was written by Michael McAuliffe and I modified the script for this experiment.
Figure 4.7: Vowel space as measured at the 25% point in the vowels.

Figure 4.8: Vowel space as measured at the 50% point in the vowels.
4.4.1 Vowel contrast

Vowel contrast refers to the spread of the vowel space, operationalized as the average vowel space (AVS) (the mean Mahalanobis distance between categories). A summary of the AVS distance results is displayed in Figure 4.10. The statistical analysis involved a number of iterations of model construction before an adequate fit was found that also met the assumptions for model validity. These steps are outlined below.

A linear mixed effects analysis was performed using the general formula outlined in Section 2.6. The two independent variables were type of oral perturbation and type of visual feedback, and the maximal random effects structure was specified. Visual inspection of the residual plots of these models revealed clear heteroscedasticity.

Since the raw data followed a gamma distribution (i.e. the values were continuous, positive, and there was right-skewing), generalized linear mixed effects models were tested, specifying that the data be fit to a gamma distribution. Models
Figure 4.10: Means and standard errors for AVS distance for each condition at three points in the vowel.

were constructed and tested using three different R packages: lme4, glmmADMB (Skaug et al., 2014), and MASS (using glmmPQL) (Venables and Ripley, 2002). There were problems with all of these models, either in terms of failure to converge, which was not resolved by simplifying the random effects structure, or heteroscedasticity and non-normality of the residuals. It is possible that these issues were due to implementation problems inherent to the packages.\footnote{The possibility of implementation problems for gamma GLMMs is noted at http://glmm.wikidot.com/faq.}

It is generally preferable to fit data to the appropriate distribution rather than transforming data and fitting them to a normal distribution. However, given the problems encountered with implementing a gamma model, data transformation was tested instead. For an underlying gamma distribution a cube root transformation is required to normalize the distribution. The cube root of each AVS distance was calculated and these values were used as the dependent variable in the models.

A linear mixed effects model was once again constructed, with the default fixed effects structure. The maximal random effects structure was used, although this differed between the models constructed for the different points in the vowels. For the models of vowel contrast at the 25% and 50% points, random effects included...
intercepts for participants and phonemes, as well as by-participant and by-phoneme random slopes for the effect of the interaction between type of oral perturbation and type of visual feedback. Visual inspection of the residual plots for these models revealed no obvious deviations from homoscedasticity.

The models for the 75% point in the vowel failed to converge with this random effects structure. To address this, the random effect for phoneme was simplified to include the intercept as well as by-phoneme random slopes for the effect of type of oral perturbation and type of visual feedback, minus the interaction term. While this model converged, visual inspection of the residual plots revealed deviations from homoscedasticity. The distribution of the cube-root AVS distances was checked to see if this could be driving the problem. The distribution still exhibited some right-skewing, suggesting that the underlying distribution was actually lognormal rather than gamma

9 Log AVS distances were calculated for this measurement point, and the linear mixed effects models re-fit with these distances as the dependent variable. The residual plots for these models were improved relative to those for the models using the cube root dependent variable, so these models were used to evaluate the effect of the experimental variables on AVS distance at the 75% point in the vowel.

10 Table 4.3 gives a summary of the fixed and random effects structures and data transformations used for each statistical model.

The models for each measurement point in the vowel is considered in turn. Model comparison at the 25% time point revealed a significant effect of visual feedback ($\chi^2(1) = 4.1681$, $p < .05$); the presence of visual feedback increased the AVS distance. The interaction at this time point was not significant. Table 4.4 shows the fixed effects coefficients, standard errors, and t-values for the model with the significant visual feedback effect.

Model comparison at the 50% time point revealed that both the effect of visual

9 The lognormal and gamma distributions are continuous probability distributions which take only positive real numbers. A simple procedure for determining which of these distributions a variable follows, is to take the logarithm of the variable; if the variable is from a lognormal distribution the logarithm will be normally distributed, while if it’s from a gamma distribution the logarithm will exhibit left-skewing.

10 While the residual plots were improved by using log AVS distances instead of cube root AVS distances, the p-values obtained from the subsequent likelihood ratio tests were very similar for both sets of models.
Table 4.3: Summary of the dependent variables and effects structures used in the statistical models of vowel contrast.

<table>
<thead>
<tr>
<th>DV</th>
<th>Fixed Effects Structure</th>
<th>Random Effects Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>25% √AVS</td>
<td>type of oral perturbation * type of visual feedback</td>
<td>(1 + oral * visual</td>
</tr>
<tr>
<td>50% √AVS</td>
<td>type of oral perturbation * type of visual feedback</td>
<td>(1 + oral * visual</td>
</tr>
<tr>
<td>75% log(AVS)</td>
<td>type of oral perturbation * type of visual feedback</td>
<td>(1 + oral * visual</td>
</tr>
</tbody>
</table>

Table 4.4: Fixed effects for the model of (cube root) AVS (25%) with the significant visual feedback effect.

feedback and the interaction between type of oral perturbation and type of visual feedback were close to significance (visual feedback effect: $\chi^2(1) = 3.5346, p = .0601$; interaction: $\chi^2(1) = 2.9687, p = .08489$). There was an overall tendency for the AVS to be increased with visual feedback, and this tendency was strengthened during normal production compared to bite block production. Table 4.5 shows the fixed effects coefficients, standard errors, and t-values for the model with the interaction.

Model comparison at the 75% time point failed to find any significant effects.
Table 4.5: Fixed effects for the model of (cube root) AVS (50%) with a near-significant interaction of the predictors.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Coefficient (Std. Error)</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.65</td>
<td>30.61</td>
</tr>
<tr>
<td>oral(biteblock)</td>
<td>−0.03</td>
<td>−0.52</td>
</tr>
<tr>
<td>visual(video)</td>
<td>0.10</td>
<td>2.71</td>
</tr>
<tr>
<td>oral(biteblock):visual(video)</td>
<td>−0.10</td>
<td>−1.73</td>
</tr>
</tbody>
</table>

4.4.2 Vowel dispersion

Vowel dispersion refers to the spread of vowel categories, operationalized as the mean of the Euclidean distances between each vowel token and the mean of the vowel category. The mean Euclidean distance results for the four conditions at each time point in the vowel are shown in Figure 4.11. The statistical analysis involved a number of iterations of model construction before an adequate fit was found that also met the assumptions for model validity. These steps are outlined here.

A linear mixed effects analysis using the default formula outlined in Section 2.6 was performed to determine the relationship between Euclidean distance and the two independent variables: type of oral perturbation and type of visual feedback. Phoneme category was also added as a fixed effect, rather than a random effect. Phonemes, rather than items, were included as a random effect in the other analyses of this experiment since item and phoneme category are conflated in the stimuli (i.e. there is only one word for each phoneme category). Additionally, only a subset of the possible English vowels were used in the experiment and thus the levels of the phoneme factor were not exhausted, which is one possible criterion for including a variable as a random effect. However, the set of possible phonemes is a small one, in contrast to the set of possible participants, for example, and sampling only a small number of factor levels from a small set can be reason to include the variable as a fixed effect. In such a situation it can also be problematic to specify the variable as a random effect, and this was the case for the analysis of Euclidean
distance. These models resulted in random effects correlations that were all at 1.0 or close to 1.0, even once the slope term was simplified, suggesting that the random effects structure was overfitting the data. To address this it was decided to model phoneme category as a fixed effect, with a by-participant random slope for phoneme. This also has the advantage of allowing any differences between vowels to be observed. Visual inspection of residual plots with this model structure revealed clear heteroscedasticity and also showed that the residuals were not normally distributed.

The procedure for fitting the data to a gamma distribution in a generalized linear mixed effects model that was outlined for the vowel contrast data above was also tested here. The same problems were encountered, so a cube root transformation of the data was performed. The cube root of each Euclidean distance was calculated and these values were used as the dependent variable in the models. New linear mixed effects models were constructed, with maximal random effects structures, although due to convergence issues this differed between the models constructed for the different points in the vowels.

In the model for the 75% point in the vowel, the random effects included intercepts for participants, as well as by-participant random slopes for the effect of

**Figure 4.11:** Means and standard errors for Euclidean distance for each condition at three points in the vowel.
Table 4.6: Summary of the dependent variables and effects structures used in the statistical models of vowel dispersion.

<table>
<thead>
<tr>
<th>DV</th>
<th>Fixed Effects Structure</th>
<th>Random Effects Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>25%</td>
<td>$\sqrt{3}AVS$ type of oral perturbation * type of visual feedback + phoneme</td>
<td>$(1 + oral + visual + phoneme</td>
</tr>
<tr>
<td>50%</td>
<td>$\sqrt{3}AVS$ type of oral perturbation * type of visual feedback + phoneme</td>
<td>$(1 + oral + visual + phoneme</td>
</tr>
<tr>
<td>75%</td>
<td>$\sqrt{3}AVS$ type of oral perturbation * type of visual feedback + phoneme</td>
<td>$(1 + oral * visual + phoneme</td>
</tr>
</tbody>
</table>

The models for each measurement point in the vowel are considered in turn. Model comparison at the 25% point revealed a significant effect of visual feedback ($\chi^2(1) = 4.7673, p = .029$), with Euclidean distance decreasing with visual feedback, and a significant effect of phoneme ($\chi^2(6) = 20.069, p < .01$), with the high back vowels /u, U/ exhibiting the greatest vowel dispersion.\(^{[11]}\) The effect of oral perturbation...
<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>(Std. Error)</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.66</td>
<td>(0.13)</td>
<td>28.47</td>
</tr>
<tr>
<td>oral (biteblock)</td>
<td>0.09</td>
<td>(0.05)</td>
<td>1.99</td>
</tr>
<tr>
<td>visual (video)</td>
<td>-0.10</td>
<td>(0.04)</td>
<td>-2.29</td>
</tr>
<tr>
<td>phoneme (æ)</td>
<td>0.12</td>
<td>(0.07)</td>
<td>1.75</td>
</tr>
<tr>
<td>phoneme (ɛ)</td>
<td>-0.04</td>
<td>(0.09)</td>
<td>-0.49</td>
</tr>
<tr>
<td>phoneme (i)</td>
<td>-0.12</td>
<td>(0.09)</td>
<td>-1.27</td>
</tr>
<tr>
<td>phoneme (i)</td>
<td>-0.23</td>
<td>(0.09)</td>
<td>-2.65</td>
</tr>
<tr>
<td>phoneme (u)</td>
<td>0.29</td>
<td>(0.09)</td>
<td>3.35</td>
</tr>
<tr>
<td>phoneme (u)</td>
<td>0.43</td>
<td>(0.12)</td>
<td>3.51</td>
</tr>
</tbody>
</table>

**Table 4.7:** Fixed effects for the model of (cube root) Euclidean distance (25%) with the significant visual feedback and phoneme effects.

turbation narrowly missed significance ($\chi^2(1) = 3.6844$, $p = .05492$). There was a tendency for the Euclidean distance to increase during bite block productions. The analysis failed to find a significant interaction at this measurement point. Table 4.7 shows the fixed effects coefficients, standard errors, and t-values for the model with the significant visual feedback and phoneme effects.

Model comparison at the 50% point revealed a significant interaction between type of oral perturbation and type of visual feedback ($\chi^2(1) = 4.185$, $p < .05$). The contrasts for the interaction were obtained using multcomp (Hothorn et al., 2008). Post-hoc comparisons revealed that Euclidean distance was smallest during normal production with visual feedback. There was a significant decrease in Euclidean distance in the NoBB-Video condition compared to the BB-NoVideo condition ($z = -2.806$, $p = .02477$ (adjusted p-value, single step method)) and the BB-Video condition ($z = -3.133$, $p < .01$ (adjusted p-value, single step method)). There was also a marginally significant decrease in Euclidean distance in the NoBB-Video condition compared to the NoBB-NoVideo condition ($z = -2.391$, $p = .07531$ (adjusted p-value, single step method)). In addition to the significant interaction, there was
Table 4.8: Fixed effects for the model of (cube root) Euclidean distance (50%) with the significant interaction between type of oral perturbation and type of visual feedback.

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>(Std. Error)</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.56</td>
<td>(0.15)</td>
<td>24.18</td>
</tr>
<tr>
<td>oral(biteblock)</td>
<td>0.07</td>
<td>(0.06)</td>
<td>1.10</td>
</tr>
<tr>
<td>visual(video)</td>
<td>-0.12</td>
<td>(0.05)</td>
<td>-2.39</td>
</tr>
<tr>
<td>phoneme(æ)</td>
<td>0.07</td>
<td>(0.08)</td>
<td>0.91</td>
</tr>
<tr>
<td>phoneme(ε)</td>
<td>-0.02</td>
<td>(0.11)</td>
<td>-0.15</td>
</tr>
<tr>
<td>phoneme(ι)</td>
<td>-0.01</td>
<td>(0.12)</td>
<td>-0.12</td>
</tr>
<tr>
<td>phoneme(ι)</td>
<td>-0.18</td>
<td>(0.11)</td>
<td>-1.61</td>
</tr>
<tr>
<td>phoneme(u)</td>
<td>0.26</td>
<td>(0.08)</td>
<td>3.46</td>
</tr>
<tr>
<td>phoneme(u)</td>
<td>0.41</td>
<td>(0.11)</td>
<td>3.70</td>
</tr>
<tr>
<td>oral(biteblock):visual(video)</td>
<td>0.13</td>
<td>(0.06)</td>
<td>2.05</td>
</tr>
</tbody>
</table>

also a significant effect of phoneme ($\chi^2(1) = 19.883, p < .01$), with the high back vowels /u, o/ exhibiting the greatest vowel dispersion. Table 4.8 shows the fixed effects coefficients, standard errors, and t-values for the model with the significant interaction.

Model comparison at the 75% point revealed a significant effect of phoneme ($\chi^2(6) = 22.728, p < .001$), with the front vowels /i, i, e/ exhibiting the least vowel dispersion. The effect of oral perturbation was close to significance ($\chi^2(1) = 3.1981, p = .07372$); there was a tendency for Euclidean distance to increase during bite block production compared to normal production. The analysis failed to find a significant interaction at this measurement point. Table 4.9 shows the fixed effects coefficients, standard errors, and t-values for the model with the significant phoneme effect. Figure 4.12 shows the differences between the vowels at each measurement point in the vowel.
<table>
<thead>
<tr>
<th>Effect</th>
<th>Coefficient</th>
<th>(Std. Error)</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.86</td>
<td>(0.13)</td>
<td>28.73</td>
</tr>
<tr>
<td>oral(biteblock)</td>
<td>0.13</td>
<td>(0.06)</td>
<td>1.93</td>
</tr>
<tr>
<td>visual(video)</td>
<td>-0.03</td>
<td>(0.04)</td>
<td>-0.64</td>
</tr>
<tr>
<td>phoneme(æ)</td>
<td>-0.24</td>
<td>(0.08)</td>
<td>-2.99</td>
</tr>
<tr>
<td>phoneme(ɛ)</td>
<td>-0.50</td>
<td>(0.09)</td>
<td>-5.79</td>
</tr>
<tr>
<td>phoneme(i)</td>
<td>-0.47</td>
<td>(0.09)</td>
<td>-4.98</td>
</tr>
<tr>
<td>phoneme(u)</td>
<td>-0.41</td>
<td>(0.12)</td>
<td>-3.45</td>
</tr>
<tr>
<td>phoneme(ʊ)</td>
<td>-0.21</td>
<td>(0.12)</td>
<td>-1.76</td>
</tr>
<tr>
<td>phoneme(u)</td>
<td>0.05</td>
<td>(0.12)</td>
<td>0.42</td>
</tr>
</tbody>
</table>

**Table 4.9:** Fixed effects for the model of (cube root) Euclidean distance (75%) with the significant effect of phoneme.
Figure 4.12: Means and standard errors for Euclidean distance for each phoneme in each condition at each measurement point in the vowel.
4.4.3 Lower face magnitude of motion

Lower face magnitude of motion (MM) was calculated from optical flow analysis of the video recordings made during the experiment. The mean MM results for the four conditions are shown in Figure 4.13. As with the acoustic measures, the statistical analysis involved a number of iterations of model construction before an adequate fit was found that also met the assumptions for model validity. These steps are outlined here.

A linear mixed effects analysis was performed using the default formula outlined in Section 2.6. The two independent variables were type of oral perturbation and type of visual feedback, and the maximal random effects structure was specified. Visual inspection of the residual plots of these models revealed clear heteroscedasticity and also showed that the residuals were not normally distributed.

The raw MM values exhibited considerable right-skewing. Unlike with the acoustic measures this was not sufficiently improved by a cube root transforma-
tion. However, a log transformation did result in a more symmetrical distribution, so these values were used as the dependent variable in the model. Inspection of this new model showed that the residuals were still somewhat heteroscedastic. To address this, outliers were identified by calculating z-scores for the untransformed MM values for each participant and excluding those that were beyond ±2.5 (approx. 2.5% of all tokens, leaving 2835 data points in the analysis).

The correlations between the fixed effects parameters for this model also suggested some degree of collinearity between the experimental variables. This was addressed by sum coding the variables instead of treatment coding (as described in Section 2.6.3). A new model was fitted and model inspection confirmed that the residuals were homoscedastic and normally distributed, and also showed reduced fixed effects correlations.

One final problem was the high random effects correlations for the phoneme random variable, suggesting that the model was overfitting the data. The significance of the slope parameters was assessed by likelihood ratio tests between models with and without the various slopes. This process revealed that, for the phoneme random effect, only a random slope for the oral perturbation was required. Subsequent models used this random effects structure for the phoneme random variable alongside the maximal random effects structure for the participant variable.

Model comparison revealed that the effect of oral perturbation was close to significance ($\chi^2(1) = 3.1591, p = .0755$), with a tendency for MM to decrease during bite block production compared to normal production. The analysis failed to find an effect of visual feedback or an interaction between oral perturbation and visual feedback. Table 4.10 shows the fixed effects coefficients, standard errors, and t-values for the model.

Closer inspection of the OFA results reveals considerable interspeaker variation. The mean MM for each vowel in each condition and for each participant is shown in Figure 4.15 and Figure 4.16, highlighting the variation across participants in response to both the bite block and visual feedback. A summary of the mean MM patterns is shown in Figure 4.14 and a description of the patterns follows.

These figures show that in response to visual feedback, for both normal productions and bite block productions, the different possibilities for mean MM changes are all realized; there are examples of increases and decreases, as well as no change.
<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.80</td>
<td>-17.25</td>
</tr>
<tr>
<td>NoBB vs. BB</td>
<td>-0.09</td>
<td>-1.84</td>
</tr>
<tr>
<td>NoVideo vs. Video</td>
<td>-0.01</td>
<td>-0.72</td>
</tr>
</tbody>
</table>

Table 4.10: Fixed effects for the model of (log) magnitude of motion. The effect of oral feedback was close to significance.

Figure 4.14: Different patterns of magnitudes of motion of the lower face in response to visual feedback. The magnitudes were pooled for each group of participants, with a separation between high vowels (/i i u/) and non-high vowels (/e æ ø/).
Five participants (P06, P07, P08, P11, and P13) exhibited quite large changes in mean MM in response to visual feedback, although participant P06’s responses patterned in a different way from the other four participants. The biggest changes were observed with the non-high vowels.

During normal production there was a tendency for MM to increase with visual feedback. Examples of this can be seen for participants P07 [æ], P08 [æ], P11 [æ], and P13 [ɛ, æ, a]. Participant P2 [o] also followed this pattern. Two notable exceptions can be seen from participants P06 [ʊ] and P12 [æ], who showed decreased MM.

During bite block production there was a tendency for MM to decrease with visual feedback. Examples of this can be seen for participant P07 [ɛ, a]. Participant P01 [ɛ] also followed this pattern. For the high vowels the changes in MM with visual feedback were more variable and of a smaller magnitude. There are high vowel examples which follow the pattern described for the non-high vowels: an MM increase with visual feedback during normal production (P13 [i]) and an MM decrease with visual feedback during bite block production (P07 [i], P08 [u]). However the opposite pattern also occurred. For example, participant P06 [o] exhibited decreased MM with visual feedback during normal production and participant P13 [o] exhibited increased MM with visual during bite block production.
Figure 4.15: Magnitudes of motion of the lower face for each participant’s production in each condition. The magnitude is shown for each of the three non-high vowels. (NoBB = normal production, BB = bite block production)
Figure 4.16: Magnitudes of motion of the lower face for each participant’s production in each condition. The magnitude is shown for each of the four high vowels. (NoBB = normal production, BB = bite block production)
Given that there were between-participant differences in lower face motion in response to visual feedback, it is possible that the enhanced acoustic contrast with visual feedback reported in Section 4.4.1 was driven by those participants who had the greatest lower face motion. Correlations between mean MM and mean AVS were calculated to investigate this possible relationship. This was not part of the planned analysis, but is presented to gain further insight into the possible effects of visual feedback. AVS distances from the 25% and 50% points in the vowels were used, as visual feedback had a significant effect at these points. These measurement points also most closely match the temporal span over which mean MM was calculated (i.e. the portion of the word up to the mid-point of the vowel). For each participant, the mean of each measure from each condition was calculated, with high and non-high vowels calculated separately in order to capture some of the differences observed in Figure 4.14. The correlation between mean MM and mean AVS was calculated for the group of participants who had an articulatory reaction to visual feedback (P07, P08, P11, P13) and the group of participants who did not have an articulatory reaction to visual feedback (P01, P02, P03, P04, P05, P09, P10, P11). Note that participant P06 was not included. While this participant did have an articulatory reaction to visual feedback, it was in the opposite direction to the other participants (MM decreased instead of increased with visual feedback during normal production). It was decided to exclude this participant rather than calculate correlations for such a small subset of the data. The data are shown in Figure 4.17.

Spearman’s correlations were calculated since the data were not normally distributed. There was a significant correlation between mean MM and mean AVS at 25% for both the no articulatory reaction group ($r_s = .37, p < .01$) and the articulatory reaction group ($r_s = .38, p < .05$). There was also a significant correlation between mean MM and mean AVS at 50% for both the no articulatory reaction group ($r_s = .42, p < .001$) and the articulatory reaction group ($r_s = .57, p < .001$). For both groups, as mean MM increased so too did the mean AVS. This suggests that the increased acoustic contrast with visual feedback observed in Section 4.4.1 was not solely due to the participants who produced greater lower face motion. In fact, Figure 4.17 suggests that it was the participants who didn’t have an articulatory reaction to visual feedback who tended to produce greater acoustic contrast.
Figure 4.17: Correlations between mean MM and mean AVS at the 25% point in the vowel (top) and between mean MM and mean AVS at the 50% point in the vowel (bottom). Within each figure the data grouped in terms of articulatory reaction to visual feedback. Dotted lines represent the median of each variable. Scatterplot smoother lines include 95% confidence intervals.
4.5 Discussion

The results showed clear effects of visual feedback on the acoustic measures, as well as more subtle effects on the articulatory measures.

Visual feedback enhanced vowel contrast at the beginning of the vowel (25% point), and there was a tendency for visual feedback to enhance contrast at the vowel midpoint during normal production more than bite block production. Visual feedback also reduced vowel dispersion. At the beginning of the vowel this occurred for both normal and bite block production. The interaction between type of oral perturbation and type of visual feedback at the vowel midpoint showed that dispersion was minimized during normal production with visual feedback in comparison to both of the bite block conditions, as well as in comparison to normal production with no visual feedback, although this latter contrast as only marginally significant. No effect of visual feedback was observed at the end of the vowel. Overall the visual feedback effect on the acoustic measures is consistent with the predictions from Section 4.2.3; visual feedback enhanced contrast and reduced dispersion. However, the initial prediction was that this effect would be greatest during bite block production, but this was not borne out in the results. Instead, there tended to be a greater effect of visual feedback during normal production.

These results complement the findings reported by Ménard and colleagues (2009; 2013) comparing blind and sighted speakers’ productions of French vowel contrasts (Section 4.2.1). These studies used a measure of average vowel space similar to the one used in the present experiment, but based on Euclidean distances rather than Mahalanobis distances, and with the inclusion of F3 information to better capture French rounding contrasts. Sighted speakers were found to produce greater acoustic contrast than blind speakers; in the present experiment, acoustic contrast was greatest when visual feedback was present compared to when it was absent. In terms of within-category vowel dispersion, [Ménard et al.] (2013) found that sighted speakers’ productions were less dispersed than blind speakers’ productions, which is also similar to the present results which found reduced vowel dispersion in the presence of visual feedback. The results from these studies and the experiment presented in this chapter support a role for visual information during speech production. Vowels produced with visual feedback are more acous-
tically contrastive relative to productions without visual feedback, just as vowels produced by speakers with normal vision are more acoustically contrastive than vowels produced by speakers with congenital visual deprivation.

The effect of visual feedback was not consistent across the measurement points during the vowels. The results from Section 4.4.1 and Section 4.4.2 have been re-plotted in Figure 4.18 to show this more clearly and to consider how the relationship between contrast and dispersion changes throughout the vowels. Vowel durations\(^\text{12}\) were also checked, and a linear mixed effects model with a maximal random effects structure showed that there was no effect of the type of visual feedback or oral perturbation on vowel duration, nor was there an interaction between the two independent variables.

In terms of vowel contrast, the AVS measure was greatest at the 25\% and 50\% points (usually slightly greater at the midpoint) and lowest at the 75\% point. A visual inspection of the plots suggests that this pattern is weakest for vowels produced without an oral perturbation (top left cell of Figure 4.18) and without visual feedback; instead the AVS remained relatively stable across the time points of the vowels. Figure 4.18 also highlights that visual feedback had a bigger effect on vowel contrast than the bite block did, as supported by the statistical analysis. In terms of vowel dispersion, the general pattern was for dispersion to be greatest at the beginning of the vowel and then decrease and stabilize by the middle of the vowel. This was the case for vowels in the BB-NoVideo, NoBB-NoVideo, and NoBB-Video conditions. Dispersion for the BB-Video condition stayed at a fairly consistent level across the vowels.

In their study looking at the role auditory feedback plays in compensating for a bite block, Lane et al. (2005) describe a complementary relationship between contrast and dispersion: “high dispersion leads to low contrast (and conversely)” (p. 1638). When the acoustic goal region is large (i.e. highly dispersed), “the trajectory, guided by least effort [...] passes through the most proximal parts of the goal regions, thereby reducing its travel and hence vowel separation” (p. 1638). The general pattern observed in the results of this experiment is for contrast to

\(^{12}\)Mean vowel durations in milliseconds for the four conditions were as follows (standard deviations in parentheses): NoBB-NoVideo 272 (54); NoBB-Video 268 (57); BB-NoVideo 265 (59); BB-Video 265 (61).
Figure 4.18: Results from Section 4.4.1 and Section 4.4.2 re-plotted to show how the relationship between vowel contrast (top) and vowel dispersion (bottom) changes throughout the vowel.

be at its maximum and dispersion to be at its minimum in the middle of vowel, which aligns with the proposed complementarity. This pattern is what one would expect for the ‘steady state’ portion of the vowel, and it was the midpoint that was also measured by Lane et al. In the present experiment, this relationship between contrast and dispersion was (roughly) maintained throughout the vowel for the condition involving normal production and no visual feedback—that is, the condition that is most similar to typical speech. However, there was a considerable decrease
in vowel contrast at the 75% measurement point for the other conditions, but this was not accompanied by a large increase in dispersion. While this measurement point was intended to primarily capture the end of the steady state portion of the vowels, it is possible that it sometimes captured formant transitions into the following consonant, especially for vowels that have more dynamic trajectories such as /æ/ and /ʌ/.

This possibility will first be considered for the contrast between the presence and absence of visual feedback during normal production (top left cell of Figure 4.18). Vowel contrast increased with visual feedback at the first two measurement points but decreased at the third measurement point. If the effects of transitioning into the following consonant counteracted the visual feedback effect then this decrease in contrast would be expected; all vowels were followed by /d/, in which case F2 for the front vowels would be expected to lower during the transition and F2 for the back vowels would be expected to raise, the net result being a reduction in contrast. One way to test this account would be to look at the complete vowel trajectory. If correct, then we would expect to see more dynamic trajectories with visual feedback compared to without visual feedback.

Comparing bite block productions (top right cell of Figure 4.18) to normal production without visual feedback, there is little difference in contrast at the first two measurement points, but a large decrease in contrast at the end of the vowels produced with a bite block. In this case it is possible that the influence of the following consonant may start sooner during bite block productions as the speaker attempts to make a constriction in the presence of the bite block, thus the contrast is even more reduced than at the comparable point in the normal productions. Bite block studies do suggest that consonants are more affected by the perturbation than vowels. For example, McFarland and Baum (1995) found that participants were unable to completely compensate for a bite block during productions of /t p k/. However, the consonants were produced in the syllable onset rather than coda, and a different type of bite block was used (one that did not protrude from the mouth). Since these studies typically take measurements at the midpoint of the vowel (e.g. Lane et al., 2005; Ménard et al., 2007) (and sometimes the first glottal pulse, e.g. McFarland and Baum, 1995), it is difficult to know how the decreased contrast at the end of vowels produced with a bite block compares to previous work.
Phoneme category was included as a fixed effect rather than random effect in the statistical analysis of vowel dispersion, due to overfitting problems. The analysis showed that phoneme was a significant factor at each measurement point in the vowel, with the most prominent pattern being that dispersion was greatest for the high back vowels /u, o/ (Figure 4.12). This is unexpected given acoustic descriptions of American English vowels (e.g., Bradlow 1995; Heald and Nusbaum 2015; Hillenbrand et al. 1995), which do not typically show the high back vowels to be substantially more variable than the other vowels. One possibility is that the high dispersion is due to some feature of the variety of American English spoken by the participants, eleven of whom were from New Mexico (mostly from Albuquerque). In this variety /u/-fronting is common and some speakers have been reported to neutralize the contrast between /u/ and /o/ before an /l/ (Labov et al. 2006). There may be more general instability in this region of the vowel space, leading to the high dispersion values reported here.

The articulatory measures revealed considerable interspeaker variation. Overall, there was no effect of visual feedback on the magnitude of motion of the lower face. For a subset of the participants, however, magnitude of motion increased with visual feedback during normal production. This effect was strongest for the non-high vowels, which is what one would expect since overall orofacial motion is much smaller for high vowels due to the low magnitude of jaw-related facial deformation. One participant produced the opposite pattern; the magnitude of motion of the lower face decreased with visual feedback during normal production, and this was observed with both non-high and high vowels. While the initial prediction that magnitude of motion would be reduced during bite block production was supported, the predicted increased in magnitude of motion with visual feedback was only partially supported.

One issue to consider is the extent to which the failure to find an effect of the experimental variables in the main analysis is due to the precision, or imprecision, of the MM measure. As discussed in Section 4.3.4.2, optical flow was calculated within an ROI that included the whole lower face. An informal comparison was made to a smaller ROI which just captured mouth movement. This showed that the pattern of movement for the two ROIs was very similar, but the magnitudes were different; the (larger) lower face ROI had a smaller magnitude of motion.
lower face ROI was chosen for ease of data extraction. While the comparison of the two ROIs confirmed that the larger ROI captured movement patterns that were representative of mouth movement, it is possible that some contrasts were obscured due to the smaller magnitudes of motion extracted from this ROI. Additionally, since a single mean value was calculated to represent lower face motion, there was a loss of information regarding temporal dynamics of the movement. Future work in this area could expand the analysis to include temporal dynamics. One way to do this would be to represent the change in summed magnitude over time with coefficients of the discrete cosine transform, an approach that has been used for modeling the temporal dynamics of vowel formants (Watson and Harrington, 1999). Applying this type of analysis to both acoustic and magnitude of motion data could provide insight into how the relationship between the two changes over time.

A first approximation of the relationship between the acoustic and articulatory results was presented in Section 4.4.3. This was a post-hoc part of the analysis presented for exploratory purposes. Positive correlations between lower face magnitude of motion and acoustic contrast were found for the group of participants who had an articulatory reaction to visual feedback as well as the group of participants who did not. As such, the enhanced acoustic contrast found with visual feedback must not be due solely to those participants who produced greater lower face motion with visual feedback. The correlations in Figure 4.17 also suggest that the group of participants who did not have an articulatory reaction to visual feedback tended to produce greater acoustic contrast compared to the participants who did have an articulatory reaction to visual feedback. Further research is required to specifically test whether speakers have a preference for responding with greater acoustic or articulatorily contrast in the presence of visual feedback. It would also be beneficial to include simultaneous recordings of tongue and lip movements to see if similar trade-offs are made in the presence of visual feedback to those reported by Ménard et al. (2013) for blind and sighted speakers.

In the present experiment, visual feedback tended to have the greatest effect on vowels during normal production rather than bite block production, contrary to the initial prediction. This may have been due to the physical limitations introduced by the bite block, by limiting the range of articulatory movement available
to speakers in such a way that the changes made in response to visual feedback during normal production were no longer possible. While speakers are able to produce extreme tongue positions with a bite block in place, lip position compensation does not always occur (e.g. for /i/: Gay et al., 1981). The bite block used in the present experiment—chopsticks that protruded from the side of the mouth—may have impeded both tongue and lip positioning more than the bite blocks used in other experiments, which are typically a small block that is held between the molars or pre-molars. However, Figure 4.15 and Figure 4.16 show that some participants produced variation in lower face magnitude of motion during bite block production depending on whether visual feedback was available. For most of these cases, the magnitudes decreased when visual feedback was available. This suggests that the bite block did not substantially limit movement, at least not for all participants. An alternative explanation for the lack of visual feedback effect during bite block production appeals to the novelty of the task. Not only did participants have to watch themselves speak, but they did so with an object protruding from the mouth. The strangeness of this may have lead to active inhibition of changes to their speech production, especially changes that were visually discernible. This issue of visual feedback differences with perturbed and unperturbed speech will be returned to in Chapter 5 in the context of possible limiting factors in multimodal feedback integration.

In this experiment visual feedback was shown to affect the production of vowels. This was particularly clear for the acoustic measures; when visual feedback was available vowel contrast increased and vowel dispersion decreased. The articulatory measures were more variable; only a subset of participants produced greater magnitudes of lower face motion when visual feedback was available. Overall, these findings are in line with the hypothesis that visual feedback can enhance speech production.
Chapter 5

Discussion

5.1 Overview

This final chapter presents a summary of the experimental findings reported in Chapter 3 and Chapter 4, and relates these to theoretical issues concerning the multimodal nature of speech production. The present research is also relevant to questions concerning the link between perception and production, and a proposal for future research in this area is outlined.

5.2 Summary of experimental results

The experiments reported in this dissertation tested speakers’ ability to incorporate visual speech feedback. While visual feedback is an atypical source of speech feedback, it is temporally compatible with auditory and somatosensory feedback since the time-varying properties of these signals are all generated by the same act of speaking. Given this compatibility, visual feedback was predicted to enhance speech output, providing reinforcement of the typical feedback signals especially during difficult speaking conditions. Real-time visual feedback, of the sort one would see when looking in a mirror, was presented during two perturbation tasks: one which delayed the auditory feedback and one which introduced an oral perturbation with a bite block.

The two experiments in Chapter 3 compared the effects of different types of
visual feedback (dynamic and static) had when paired with normal and delayed auditory feedback. The different visual feedback conditions were randomized within a block in order to make it less likely that participants would actively use the visual feedback as a strategy to counteract the disruptive effects of delayed auditory feedback (DAF). On the basis of previous research, dynamic visual feedback was predicted to enhance the production of whole utterances when paired with DAF, decreasing utterance duration and the number of speech errors. The results showed an increase in utterance duration with dynamic visual feedback but no significant change in the number of speech errors, suggesting that sustained exposure to visual feedback is required before speech enhancement is observed.

The second experiment in Chapter 3 presented the different types of visual feedback (dynamic, static, no visual feedback) in consistent blocks. In doing so, it tested whether the predicted visual feedback effects would be observed when visual feedback was presented consistently, and it also allowed a more careful consideration of the properties of visual feedback that may drive speech production changes. As in the first experiment, utterance duration increased with dynamic visual feedback. One possible explanation for this effect—which was in the opposite direction to the initial prediction—is that durational increases over specific segments contributed to the overall durational increases. Previous research has reported that the closed portion of the syllable (i.e. the period between the movement offset of the VC gesture and movement onset of the CV gesture) are lengthened in response to DAF, particularly for visible places of articulation (i.e. labial, labiodental); this effect could be reinforced by dynamic visual feedback, resulting in even longer durations. Unlike the first experiment, a small reduction in speech errors was observed; there was an overall reduction in speech errors with dynamic visual feedback, but this was only significant for those participants who were minimally disrupted by DAF. Interestingly, this effect occurred when dynamic visual feedback was paired with normal, but not delayed, auditory feedback. This finding suggests that temporal cohesion among multimodal feedback signals may be important in speech production. There was no effect of static visual feedback on either of the measures, suggesting that it is instead the time-varying properties of visual feedback which are important for enhancing speech output.

Chapter 4 presented the results of a bite block experiment, comparing the ef-
fects of the presence versus absence of visual feedback when speech was produced with a bite block in place or without any oral obstructions. Participants produced /hVd/ words (heed, hid, head, had, hod, hood, who’d) in each condition, and changes in acoustic contrast and dispersion, as well as lower face magnitude of motion, were measured. Acoustic speech output was enhanced in the predicted direction: vowel contrast increased and vowel dispersion decreased during productions with visual feedback. This effect was greatest at the beginning of vowels and tended to be stronger during productions without the bite block. There was considerable inter-speaker variation for the lower face motion results. A subset of the participants tended to produce greater magnitudes of motion with visual feedback, especially for the non-high vowels. In addition to a modest positive correlation between lower face motion and acoustic contrast, it was also noted that the subset of participants who produced greater magnitudes of motion with visual feedback tended to produce less acoustic contrast than the participants who produced smaller magnitudes of motion.

Overall these results support the hypothesis that visual feedback can enhance speech production. The fluency enhancement observed in Experiment 2 of the DAF experiments was more limited (i.e. it was only significant when additional predictors were added to the statistical model, and only for normal auditory feedback) than the acoustic contrast enhancement observed in the bite block experiment. This may be due in part to the different stimuli used in the experiments: participants repeated whole sentences in the DAF experiments and monosyllabic words in the bite block experiment. Limiting speech production to simple strings, such as a CVC syllable, instead of structurally complex sentences, can make it easier to observe the effects of experimental manipulations since there are fewer linguistic factors muddying the waters, so to speak. However, data from productions of whole utterances provide an important reminder that speech is a highly complex process, and in this context the effects of experimental manipulations will often play out in more subtle, and potentially even different, ways. For example, in the second DAF experiment, the visual feedback both enhanced, in terms of speech error reduction, and hampered, in terms of durational increases, speech production.

It is true that the analyses presented in this dissertation show that the effect visual feedback has on speech production is, overall, quite small. This is perhaps un-
surprising given that visual feedback is not typically available during speech. However, the results are reassuringly consistent with a general pattern that is emerging across different populations and methodological frameworks; namely, that visual speech information enhances not only perception, but also production. Visually stimulated speech production is more accurate (Reisberg et al., 1987; Scarbel et al., 2014), and more fluent in the case of aphasic patients (Fridriksson et al., 2015, 2012) and people who stutter (Kalinowski et al., 2000). When presented as feedback of one’s own productions, visual speech information improves fluency for stutterers (Snyder et al., 2009) and non-stutterers (Jones and Striemer, 2007). Similar fluency enhancing results were reported in Chapter 3, and the new finding that visual feedback also increases vowel contrast and reduces vowel dispersion was reported in Chapter 4. There is also evidence that when speakers are deprived of visual speech information during development, due to congenital blindness, the opposite patterns are found; in comparison to sighted speakers productions’ vowel contrast is reduced and vowel dispersion is increased (Ménard et al., 2009, 2013). In addition, visible articulations exhibit smaller movements and non-visible articulations exhibit larger movements for blind compared to sighted speakers (Ménard et al., 2013). Finally, a recent brain imaging study shows that visually stimulated production recruits additional neural pathways beyond the typical auditory-motor pathway, providing a possible neural basis for these behavioral changes (Venezia et al., 2016).

Visual feedback can influence speech production, and the results presented in this dissertation have implications for theoretical issues in speech production research. Two such issues—targets of production and multimodal feedback integration—are addressed in the next section.

5.3 Multimodal speech production

5.3.1 Targets of production

When a speech sound is produced, the speaker is presumed to be attempting to achieve a target, which is typically represented in some task space or coordinate frame. Research tends to have the goal of placing the target either in articula-
tory space (e.g. Browman and Goldstein, 1992) or in auditory/acoustic space (e.g. Stevens, 1989). Compensation for altered acoustic feedback (e.g. Houde and Jordan, 1998) and the trading relations between different articulations with the same acoustic consequences (e.g. Guenther et al., 1999) are argued to be evidence for auditory/acoustic targets. Compensation for altered somatosensory feedback (e.g. Tremblay et al., 2003) and articulators functioning as coordinative structures (e.g. Kelso et al., 1984) are argued to be evidence for articulatory space. However, given the complexity of the speech processing system, it is unlikely that placing targets in only one of these task spaces could adequately capture the whole system. Models such as the perception for action control theory (PACT) (Schwartz et al., 2012) propose a middle ground between these two task spaces and aim to codify the relation between perception and production; in PACT, speech targets are described as “perceptually-shaped gestures.” This move towards more flexible models of speech production raises the question of whether there is a visual component to speech targets.

Recent work with aphasic and non-clinical populations provides support for visual speech targets and suggests that these targets may also involve dedicated neural pathways in the speech motor control architecture. Aphasic patients demonstrate increased fluency when speaking in time with an audiovisually presented recording of a speaker, compared to an audio-only presentation (Fridriksson et al., 2012). This effect, referred to as speech entrainment, is observed in those patients who still have cortical motor areas and auditory-motor interface areas relatively intact (Fridriksson et al., 2015). The existence of a neural route for mapping visual targets to motor programs was proposed as an alternative to the auditory-motor pathway, based on the fact that enhanced fluency was only observed during audiovisual speech entrainment. Venezia et al. (2016) provided evidence for this pathway in healthy speakers. As described in Section 1.1.3 a comparison was made between the covert production of CV syllable strings that were produced in response to either audio-only, visual-only, or audiovisual stimuli. Neural regions involved in audio-only stimulated productions were activated to a greater extent by visual-only and audiovisually stimulated productions, and these latter productions also involved additional sensorimotor brain regions.

Of the additional regions activated, Venezia et al. (2016) note that the left pos-
terior middle temporal gyrus is of particular interest. Not only did this region activate during covert rehearsal following visual and audiovisual stimuli, activation was also found during passive perception, especially for visual and audiovisual inputs. They propose that this region may be the site of “visual speech targets for production (i.e., high-level sensory representations of visual speech gestures)” (p. 204).

Visual speech gestures (specifically, the magnitude of motion of lower face) were measured in the bite block experiment in Chapter 4. On the face of it, the results from this experiment offer only limited support for visual targets; speech produced with visual feedback only resulted in changes to the magnitude of lower face motion for five of the thirteen participants. Of these five, four participants tended to produce non-high vowels with greater magnitudes when visual feedback was available, and one participant produced vowels with smaller magnitudes when visual feedback was available. But the acoustic and articulatory results as a whole raise a number of issues that would need to be taken into account in a model of speech motor control that combines speech targets from multiple task space, as proposed by (Venezia et al., 2016, p. 197):

“To be specific, we assume that the noted behavioral increases in speech output during or following exposure to audiovisual speech reflect the activation of a complementary set of visual speech targets (i.e., the visual patterns a talker is trying to produce) that combine with auditory speech targets to facilitate speech motor control.”

This issue is addressed in the next section in the context of multimodal feedback integration.

### 5.3.2 Multimodal feedback integration

The response to feedback and feedback manipulations has been used as evidence for the task space of targets, as described above. Feedback can not only be used to detect and correct perturbations, it can also be used to update forward models of motor control (Hickok, 2014). Venezia et al. (2016) suggest that a dedicated visuo-motor pathway for speech motor control could be relevant for feedback: “One possibility is that feedback from visual speech is used to tune internal vocal tract
control circuits in a similar fashion to auditory speech.” (p. 205). In this section, this proposal for visual feedback and its integration with other modalities is explored in the context of Hickok and colleagues (Hickok, 2012, 2014; Hickok et al., 2011) State Feedback Control (SFC) model of speech processing. While Venezia et al. don’t specifically situate their proposal for visual targets within this model, the assumptions of this model are implicit in their discussion. Implications for this proposal are discussed in light of the results presented in this dissertation, specifically the fact that the effects of visual feedback tended to be most clearly seen during unperturbed speech.

The SFC model integrates theoretical perspectives and experimental findings from speech perception, speech production, psycholinguistics, and clinical domains, with the aim of marrying auditory- and motor-centric models. The (Hierarchical) State Feedback Control (SFC) model (Hickok, 2012, 2014; Hickok et al., 2011) is an elaboration of previous work (e.g. Hickok and Poeppel’s (2004) dual stream hypothesis, and Ventura et al.’s (2009) evidence for the role of efference copies in speech motor control), providing a detailed proposal for the sensory-motor integration processes of the dorsal stream (Hickok and Poeppel, 2004).

In the SFC model, speech production begins with an intention to speak, which provides parallel input to motor and auditory phonological systems. This input activates two types of representation: a motor plan and associated sensory targets. System output comes from an articulatory controller that generates motor commands for the vocal tract, as well as sending a copy of the motor commands to an internal model of the vocal tract. This internal model is an estimate of the state of the vocal tract. The estimate is then transformed into a prediction of the sensory consequences of the motor command. The prediction is involved in two subsequent functions: fast internal monitoring and slow external monitoring. Internal monitoring considers whether the motor command that has been initiated will have the intended sensory consequences, while external monitoring considers whether the actual sensory consequences match the predicted sensory consequences. Error signals can be generated by either of these monitoring loops, providing corrective feedback to the motor controller via the internal model.

The most recent descriptions of this model include this control system at two levels: the higher level involves auditory targets at the syllable level and the lower
level involves somatosensory targets at the phoneme level (Hierarchical State Feedback Control (HSFC) Hickok, 2012, 2014). However, this division is not absolute, as phoneme and syllable targets may be distributed across the two levels, and have different weightings, depending on the phoneme or syllable. For example, sibilants have clear auditory and somatosensory targets. These two levels are presumed to interact, with one possible function of this interaction being to fine tune forward predictions generated at one level with information from the other level. For example, Hickok (2012) suggests that information about the articulatory phase from the somatosensory level might enable the auditory prediction to be more accurate. The results of this dissertation suggest that visual feedback may have some influence on these different levels of the control system. For example, visual feedback enhanced acoustic vowel contrast and reduced acoustic vowel dispersion (Chapter 4); this could be interpreted as visual feedback contributing to the production of more accurate acoustic/auditory targets. How might this influence of visual feedback be implemented?

Venezia et al. (2016) propose that the sensorimotor integration of visual speech involves separate auditory-motor and visuo-motor speech pathways, and thus complementary sets of targets across the different modalities. This proposal is consistent with their finding that neural regions in addition to the auditory-motor network were activated in response to visually stimulated production, and also findings that aphasic patients with damage to their auditory-motor network (restricted to the inferior frontal gyrus) show improved fluency when shadowing audiovisual speech compared to audio-only speech (Fridriksson et al., 2015, 2012). Venezia et al. consider, and reject, two other mechanisms for integrating visual and auditory speech signals; these are in line with proposals for visual speech integration which involve visual information increasing activation of the auditory-motor pathway (e.g. Calvert et al., 2000). The first possibility is that auditory and visual speech are first integrated, and this integrated representation serves as input to the dorsal stream, which is involved in the sensory-motor integration processes described in the SFC model. The second possibility is that visual speech is integrated directly in the dorsal stream (i.e. in the sensory-motor network rather than in the sensory system). These two possibilities are considered unlikely based on Venezia et al.’s results; while some brain regions responded to all types of input (audio-only, visual-only,
audiovisual; although most strongly for the latter two), there were regions that showed greater activation during rehearsal when stimulated by visual-only or audiovisual signals, but not audio-only inputs. These regions included the bilateral pre-central sulci and left central sulcus, the caudate nucleus, the inferior frontal gyrus, and the middle temporal gyrus.

Recall that while Venezia et al.’s (2016) proposal was based on results from visually stimulated speech production (specifically, the covert repetition of (audio)visual presentations of another speaker producing CV syllable strings), they suggest that the proposed visuo-motor pathway could be relevant for visual speech as feedback. The results from this dissertation do provide further evidence of “behavioral increases in speech output during or following exposure to audiovisual speech” (Venezia et al., 2016, p. 197); the presence of visual feedback was associated with modest improvements in fluency (Chapter 3) and enhanced vowel contrasts (Chapter 4). One interesting aspect of these results is that the visual feedback effects were most often observed during unperturbed speech; that is, when auditory feedback was not delayed and there was no bite block. This has implications for the coordination of different feedback signals, in terms of temporal alignment requirements and the relative weight of feedback.

In the second DAF experiment (Section 3.4) the predicted reduction in speech errors occurred, but only when dynamic visual feedback was paired with normal auditory feedback, and not DAF as predicted. A possible explanation for this is that temporal cohesion among the different types of feedback may be necessary in order for speech output to be enhanced.

A number of studies suggest that there is sensitivity to timing information during the process of detecting mismatches between predicted and reafferent feedback. The act of producing speech has been shown to result in a suppressed auditory cortical response, which is hypothesized to be due to an accurate match between the predicted and actual sensory consequences of speaking (Heinks-Maldonado et al., 2006). The production of faster, more rhythmically complex vowel sequences results in less suppression than single sustained vowels, indicative of a poorer match between the predicted and reafferent feedback, possibly due to timing discrepancies: “the auditory feedback predictions became more dynamic and more difficult to keep in temporal registry with the incoming auditory feedback” (Ventura et al.,
Behroozmand et al. (2011) confirmed that auditory-evoked responses are sensitive to timing mismatches caused by altering pitch feedback at different onset delays. Error detection was enhanced after the immediate onset of production, perhaps to accommodate delays in sensory feedback transmission. Later work showed that ERP responses are enhanced (i.e. the mismatch detection is more sensitive) when pitch perturbations occur earlier in the vocalization (Behroozmand et al., 2016). Sensitivity to variability in the production of vowels is also seen at the onset of vocalization; within the first 50 ms of the vowel, productions that are acoustically farther from the center of the vowel category’s distribution result in less suppression of the auditory-evoked response (i.e. an error response) (Niziolek et al., 2013). This error response is also correlated with a “corrective” change; these peripheral vowels become acoustically closer to the center of the vowel distribution by the middle of the vowel.

This work suggests that prediction errors are sensitive to timing mismatches, particularly at the beginning of an utterance and when more complex speech is used. While this work has only considered auditory feedback, it is likely that it would apply to other speech feedback channels, especially given that there is evidence for these temporal patterns of suppression response in other domains. For example, while self-produced touches result in cortical suppression in the cerebellum and reports of diminished ‘tickliness’, once a delay is introduced there is less suppression and participants progressively rate the touch as more ‘tickly’ (Blakemore et al., 2001, 2000). In perception, visual speech can be thought of as an anchor that improves a perceiver’s ability to align to a multisensory signal (Vatikiotis-Bateson and Munhall, 2015). In the context of feedback, the presence of multiple signals may also have an anchoring effect stemming from the fact that all sources of feedback are generated from the same event of speaking. The combination of this effect and the sensitivity to timing mismatches between predicted and reafferent feedback within a given feedback modality may lead to less tolerance of timing mismatches between multiple feedback signals. In the context of the HSFC model, this temporal alignment would be important in order for feedback from the different levels to be able to fine-tune predictions. In the context of visual feedback, it is the time-varying properties of the signal that establish its compatibility with other sources of feedback, adding a further need for temporal cohesion.
The results from the bite block experiment (Chapter 4) also showed that visual feedback tended to have a stronger effect—in terms of enhanced acoustic contrast and diminished vowel dispersion—on unperturbed speech. The possibility that this was due to physical limitations introduced by the bite block was discussed in Section 4.5 although this was possibility was considered unlikely. An alternative account is presented here.

In the Venezia et al. (2016) model there are complementary sets of auditory and visual targets, and the visual targets are more strongly activated when speech is produced in the context of visual stimulation. The bite block results suggest that there is more that needs to be taken into account; while visual targets may have been activated with visual feedback, when the bite block was in place the speech enhancing effects of visual feedback were diminished. One way to think of this is in terms of different weightings; somatosensory feedback became more heavily weighted than visual feedback when there was an oral perturbation to counteract. Given that one rarely sees one’s mouth, it is perhaps unsurprising that somatosensory feedback is prioritized in this particular context. (Haggard and de Boer, 2014, p. 470) compare this lack of oral visual experience to the contributions made by somatosensory and visual feedback during manual tasks: “The somatosensory innervation of the hand, although very rich, normally remains subservient to vision. [...] In contrast, within the mouth, somatosensation rules.”

However, the more extensive experience with somatosensory feedback isn’t enough to block the effects of visual feedback. And in the context of speech, it is not clear that any one feedback modality can be said to “rule” the others. Recent work suggests that individuals have modality preferences, for both feedback received during production (e.g. Lametti et al., 2012) and speech signals during perception (e.g. Gick et al., 2008). For example, using a perturbation paradigm, Lametti et al. (2012) simultaneously manipulated both auditory feedback (downward shift of F1) and somatosensory feedback (mechanical altering of jaw displacement). A negative correlation between the amount of compensation for each perturbation was found; participants who compensated more for the somatosensory perturbation compensated less for the acoustic perturbation. Some participants were also observed to only compensate for one type of perturbation. Related to this are the results from Chapter 3 and from Jones and Striener (2007) showing
that the visual feedback effect depends on how disrupted a person is by DAF (i.e. to what degree auditory feedback is preferred). Additionally, the correlations between lower face magnitude of motion and acoustic contrast in Chapter 4 suggest that some speakers respond to visual feedback by enhancing acoustic contrast and others respond by enhancing part of their articulations; these different responses could be driven by an individual’s modality preference.

5.3.3 Summary

Contrary to the initial predictions of this dissertation research, visual feedback tended to have a stronger effect on unperturbed speech than perturbed speech. While it is the case that speakers can use visual information in difficult listening conditions (e.g. Navarra and Soto-Faraco, 2007; Sumby and Pollack, 1954) and speaking conditions (e.g. Fridriksson et al., 2012; Kalinowski et al., 2000), there are additional factors that need to be considered for the integration of feedback signals during production. The demands of counteracting an oral perturbation may have the effect of increasing the weighting of somatosensory feedback, thus reducing visual feedback effects. Delaying auditory feedback may interfere with temporal cohesion among the feedback signals, with timing errors overriding visual feedback effects. In a model of speech processing such as Hickok’s (2012; 2014) HSFC model, in which there would be forward control systems for auditory, somatosensory, and visual feedback, these factors are relevant to the manner in which the different levels would interact. More broadly, the implications of these results for speech targets is that domains of representation need not be rigidly defined. There is evidence for both sensorimotor and auditory/acoustic targets, and growing evidence for visual targets as well.

5.4 Future work

This research into the effects of visual feedback on production is related to more general questions concerning the relation between production and perception. This relation can be between a speaker and a listener; for example, Pickering and Garrod’s (2004) interactive alignment account relies on a tight coupling between production and perception to facilitate dialogue. But often this relation is conceptual-
ized as being relevant to individual speaker-listeners, as in motor theories of speech perception, which hypothesize that what is being perceived is speech gestures (e.g. Liberman and Mattingly, 1985). Perkell and colleagues (2004a, 2004b) assume the opposite hypothesis; namely, that perception drives production. They investigated this issue by looking at the relationship between speakers’ perceptual acuity for vowel and sibilant contrasts and their production distinctness for these contrasts. Those speakers who had higher perceptual discrimination scores were found to also produce greater contrasts, measured in terms of tongue body movements and F1xF2 acoustic space for the vowels, and in terms of linguo-dental contact in the sublingual cavity and acoustic center of gravity for the sibilants.

Two studies involving visual information complicate this account. In a vowel discrimination task, Ménard et al. (2009) found that blind participants had higher discrimination scores than sighted participants for vowel contrasts, but sighted speakers produced a more contrastive vowel space than blind speakers. Given Perkell et al. (2004a) proposal, one would expect the blind speakers’ vowel space to be more contrastive in response to their greater perceptual acuity. Ménard et al. (2009) suggest that the effects of visual deprivation on production override the effects of perceptual acuity. Alternatively, it’s possible that blind speakers do have greater perceptual acuity, but along dimensions not tested by the experimental manipulations.

Findings from a study comparing speech-reading abilities of oneself versus others also raise questions about Perkell et al.’s (2004a) account. In line with research showing sensitivity to self-generated biological motion (e.g. Knoblich and Flach, 2001), Tye-Murray et al. (2013) demonstrated that perceivers are better at speech-reading videos of their own speech compared to the speech of others. Participants were pre-recorded reading a large list of sentences, a subset of which was then used as stimuli in a speech-reading test. They were significantly better at speech-reading themselves than other people, independent of their general lipreading ability. As part of the analysis, Tye-Murray et al. (2013) noted that there was no correlation between participants’ general speech-reading ability and how well they were speech-read by other participants. If perception drives production, we would expect to see a positive correlation here: the better a person is at speech-reading (due to high perceptual acuity), the better they are at being speech-read by
others (due to high production distinctness). Admittedly, the goal of Tye-Murray et al.’s study was not to test this perception-production link, so it is possible that the finding is due simply to the particular experimental set up which was testing another question. However, this finding, along with that from Ménard et al. (2009), suggests that there are more factors to be considered in describing the perception-production link in the context of audiovisual speech processing.

Thus the general question for future research concerns whether there is a relation between how sensitive a perceiver is to visual speech information and how visually contrastive they are as a speaker. The fact that the results reported in Chapter 4 showed that some speakers produced greater magnitudes of lower face motion in response to visual feedback suggests that there may be a positive correlation between visual perception and production, at least for some speakers.

The investigation of visual feedback can contribute a new perspective on old questions like the relation between production and perception. Exploring the relation between speaker and listener in the visual domain, in addition to the auditory and articulatory domains, is particularly germane to this question given the highly visible nature of much face-to-face communication. But visual information is also relevant to the link between production and perception for individual speaker-listeners. Not only is there well established work showing the importance of visual speech information for perception, but there is a growing body of research, of which this dissertation is a part, showing that visual information, including visual feedback of oneself, can also affect speech production.
Bibliography


148


Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. → pages 23


### Appendix A

### Stimuli

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Accepted Substitution</th>
</tr>
</thead>
<tbody>
<tr>
<td>A cup of sugar makes sweet fudge</td>
<td>“made” for “makes”</td>
</tr>
<tr>
<td>The doorknob was made of bright clean brass</td>
<td>“clear” for “clean”, “green” for “clean”, “glass” for “brass”</td>
</tr>
<tr>
<td>We need grain to keep our mules healthy</td>
<td>“meals” for “mules”</td>
</tr>
<tr>
<td>The plush chair leaned against the wall</td>
<td>“pushed” for “plush”</td>
</tr>
<tr>
<td>Bathe and relax in the cool green grass</td>
<td>“cold” for “cool”, “glass” for “grass”</td>
</tr>
<tr>
<td>Take two shares as a fair profit</td>
<td>“pick” for “take”</td>
</tr>
<tr>
<td>North winds bring colds and fevers</td>
<td>“cold” for “colds”</td>
</tr>
<tr>
<td>A gray mare walked before the colt</td>
<td>“great” for “gray”, “cold” for “colt”</td>
</tr>
</tbody>
</table>

Table A.1 continued on next page
<table>
<thead>
<tr>
<th>Sentence</th>
<th>Accepted Substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cap the jar with a tight brass cover</td>
<td>“glass” for “brass”, “bright” for “tight”, and “tap” for “cap”</td>
</tr>
<tr>
<td>The odor of spring makes young hearts jump</td>
<td>“order” for “odor”</td>
</tr>
<tr>
<td>They sliced the sausage thin with a knife</td>
<td></td>
</tr>
<tr>
<td>Take the winding path to reach the lake</td>
<td></td>
</tr>
<tr>
<td>Note closely the size of the gas tank</td>
<td></td>
</tr>
<tr>
<td>Wipe the grease off his dirty face</td>
<td></td>
</tr>
<tr>
<td>Mend the coat before you go out</td>
<td></td>
</tr>
<tr>
<td>The stray cat gave birth to kittens</td>
<td></td>
</tr>
<tr>
<td>The young girl gave no clear response</td>
<td></td>
</tr>
<tr>
<td>The meal was cooked before the bell rang</td>
<td></td>
</tr>
<tr>
<td>The frosty air passed through the coat</td>
<td></td>
</tr>
<tr>
<td>A saw is a tool used for making boards</td>
<td></td>
</tr>
<tr>
<td>The wagon moved on well oiled wheels</td>
<td></td>
</tr>
<tr>
<td>March the soldiers past the next hill</td>
<td></td>
</tr>
<tr>
<td>Place a rosebush near the porch steps</td>
<td></td>
</tr>
<tr>
<td>Both lost their lives in the raging storm</td>
<td></td>
</tr>
<tr>
<td>The cement had dried when he moved it</td>
<td></td>
</tr>
</tbody>
</table>
Table A.1 – continued from previous page

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Accepted Substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>The fly made its way along the wall</td>
<td></td>
</tr>
<tr>
<td>Do that with a wooden stick</td>
<td></td>
</tr>
<tr>
<td>Live wires should be kept covered</td>
<td></td>
</tr>
<tr>
<td>The large house had hot water taps</td>
<td></td>
</tr>
<tr>
<td>It is hard to erase blue or red ink</td>
<td></td>
</tr>
<tr>
<td>The wreck occurred by the bank on Main Street</td>
<td></td>
</tr>
<tr>
<td>Fill the ink jar with sticky glue</td>
<td></td>
</tr>
<tr>
<td>Pack the records in a neat thin case</td>
<td></td>
</tr>
<tr>
<td>That move means the game is over</td>
<td></td>
</tr>
<tr>
<td>Glass will clink when struck by metal</td>
<td></td>
</tr>
<tr>
<td>It takes a lot of help to finish these</td>
<td></td>
</tr>
<tr>
<td>Mark the spot with a sign painted red</td>
<td></td>
</tr>
<tr>
<td>The fur of cats goes by many names</td>
<td></td>
</tr>
<tr>
<td>He asks no person to vouch for him</td>
<td></td>
</tr>
<tr>
<td>Go now and come here later</td>
<td></td>
</tr>
<tr>
<td>Soap can wash most dirt away</td>
<td></td>
</tr>
<tr>
<td>The bloom of the rose lasts a few days</td>
<td></td>
</tr>
<tr>
<td>Bottles hold four kinds of rum</td>
<td></td>
</tr>
</tbody>
</table>

Table A.1 continued on next page
<table>
<thead>
<tr>
<th>Sentence</th>
<th>Accepted Substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>He wheeled the bike past the winding road</td>
<td></td>
</tr>
<tr>
<td>Drop the ashes on the worn old rug</td>
<td></td>
</tr>
<tr>
<td>The desk and both chairs were painted tan</td>
<td></td>
</tr>
<tr>
<td>The way to save money is not to spend much</td>
<td></td>
</tr>
<tr>
<td>Shut the hatch before the waves push it in</td>
<td></td>
</tr>
<tr>
<td>Crack the walnut with your sharp side teeth</td>
<td></td>
</tr>
<tr>
<td>He offered proof in the form of a large chart</td>
<td></td>
</tr>
<tr>
<td>Send the stuff in a thick paper bag</td>
<td></td>
</tr>
<tr>
<td>A quart of milk is water for the most part</td>
<td></td>
</tr>
<tr>
<td>They told wild tales to frighten him</td>
<td></td>
</tr>
<tr>
<td>The three story house was built of stone</td>
<td></td>
</tr>
<tr>
<td>The poor boy missed the boat again</td>
<td></td>
</tr>
<tr>
<td>Be sure to set the lamp firmly in the hole</td>
<td></td>
</tr>
<tr>
<td>Pick a card and slip it under the pack</td>
<td></td>
</tr>
<tr>
<td>The first part of the plan needs changing</td>
<td></td>
</tr>
<tr>
<td>The mail comes in three batches per day</td>
<td></td>
</tr>
<tr>
<td>You cannot brew tea in a cold pot</td>
<td></td>
</tr>
<tr>
<td>* The crooked maze failed to fool the mouse</td>
<td></td>
</tr>
<tr>
<td>* A sash of gold silk will trim her dress</td>
<td></td>
</tr>
</tbody>
</table>

Table A.1 – continued from previous page
Table A.1 – continued from previous page

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Accepted Substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>* Breakfast buns are fine with a hot drink</td>
<td></td>
</tr>
<tr>
<td>* Throw out the used paper cup and plate</td>
<td></td>
</tr>
<tr>
<td>^ The crunch of feet in the snow was the only sound</td>
<td></td>
</tr>
<tr>
<td>^ In the rear of the ground floor was a large passage</td>
<td></td>
</tr>
<tr>
<td>^ The man wore a feather in his felt hat</td>
<td></td>
</tr>
<tr>
<td>^ Hang tinsel from both branches</td>
<td></td>
</tr>
<tr>
<td>^ A round mat will cover the dull spot</td>
<td></td>
</tr>
<tr>
<td>^ Boards will warp unless kept dry</td>
<td></td>
</tr>
</tbody>
</table>

**Table A.1:** Stimuli used in the DAF experiments (Chapter 3). Alternate words which were accepted as substitutions during speech error coding are listed. (Sentences marked with an asterisk were not used in Experiment 2. Sentences marked with a caret were practice sentence.)

<table>
<thead>
<tr>
<th>Word</th>
<th>Example Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>heed</td>
<td>The girl did not heed their warning.</td>
</tr>
<tr>
<td>she</td>
<td>She likes to go bowling.</td>
</tr>
<tr>
<td>Word</td>
<td>Example Sentence</td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>hod</td>
<td>A hod is a tool used by builders.</td>
</tr>
<tr>
<td>sort</td>
<td>They need to sort the blocks.</td>
</tr>
<tr>
<td>who’d</td>
<td>Who’d have thought he’d win?</td>
</tr>
<tr>
<td>sheet</td>
<td>I put a fresh sheet on the bed.</td>
</tr>
<tr>
<td>hid</td>
<td>The boy hid from the bullies.</td>
</tr>
<tr>
<td>sore</td>
<td>My legs are sore after exercising.</td>
</tr>
<tr>
<td>hood</td>
<td>The hood of the car was hot.</td>
</tr>
<tr>
<td>seat</td>
<td>Take a seat by the window.</td>
</tr>
<tr>
<td>had</td>
<td>I wish I had some money.</td>
</tr>
<tr>
<td>shore</td>
<td>We walked along the shore.</td>
</tr>
<tr>
<td>head</td>
<td>His head was blocking my view.</td>
</tr>
<tr>
<td>see</td>
<td>Can you see the balloon?</td>
</tr>
<tr>
<td>short</td>
<td>The man was too short.</td>
</tr>
<tr>
<td>‹ fought</td>
<td>The soldiers fought in the war.</td>
</tr>
<tr>
<td>‹ gas</td>
<td>We need gas for the car.</td>
</tr>
<tr>
<td>‹ heard</td>
<td>I heard you talking.</td>
</tr>
<tr>
<td>‹ hide</td>
<td>The boy will hide the chocolate.</td>
</tr>
<tr>
<td>‹ how’d</td>
<td>How’d you do in the exam?</td>
</tr>
<tr>
<td>‹ leash</td>
<td>Put the leash on the dog.</td>
</tr>
</tbody>
</table>
Table A.2: Stimuli used in the bite block experiment (Chapter 4). The word list and example sentences were shown to participants prior to the experiment. (Words marked with a caret were practice words.)

<table>
<thead>
<tr>
<th>Word</th>
<th>Example Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>mat</td>
<td>Please wipe your feet on the mat.</td>
</tr>
<tr>
<td>wet</td>
<td>We got wet during the storm.</td>
</tr>
<tr>
<td>hayed</td>
<td>The grass was hayed to make cattle feed.</td>
</tr>
<tr>
<td>hoyed</td>
<td>A passing stranger hoyed me.</td>
</tr>
</tbody>
</table>