Speech motor control variables in the production of voicing contrasts and emphatic accent

Timothy Ian Pandachuck Mills
BA Hon, MSc

A thesis submitted in fulfilment of requirements for the degree of Doctor of Philosophy

to

The School of Philosophy, Psychology, and Language Sciences
College of Humanities and Social Sciences
University of Edinburgh

August 2009

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Timothy Ian Pandachuck Mills

Abstract

This dissertation looks at motor control in speech production. Two specific questions emerging from the speech motor control literature are studied: the question of articulatory versus acoustic motor control targets, and the question of whether prosodic linguistic variables are controlled in the same way as segmental linguistic variables.

In the first study, I test the utility of whispered speech as a tool for addressing the question of articulatory or acoustic motor control targets. Research has been done probing both sides of this question. The case for articulatory specifications is developed in depth in the Articulatory Phonology framework of Haskins researchers (eg Browman & Goldstein 2000), based on the task-dynamic model of control presented by Saltzman & Kelso (1987). The case for acoustic specifications is developed in the work of Perkell and others (eg Perkell, Matthies, Svirsky & Jordan 1993, Guenther, Espy-Wilson, Boyce, Matthies, Zandipour & Perkell 1999, Perkell, Guenther, Lane, Matthies, Perrier, Vick, Wilhelms-Tricarico & Zandipour 2000). It has also been suggested that some productions are governed by articulatory targets while others are governed by acoustic targets (Ladefoged 2005).

This study involves two experiments. In the first, I make endoscopic video recordings of the larynx during the production of phonological voicing contrasts in normal and whispered speech. I discovered that the glottal aperture differences between voiced obstruents (ie, /d/) and voiceless obstruents (ie, /t/) in normal speech was preserved in whispered speech. Of particular interest was the observation that phonologically voiced obstruents tended to exhibit a narrower glottal aperture in whisper than vowels, which are also phonologically voiced. This suggests that the motor control targets of voicing is different for vowels than for voiced obstruents. A perceptual experiment on the speech material elicited in the endoscopic recordings elicited judgments to see whether listeners could discriminate phonological voicing in whisper, in the absence of non-laryngeal cues such as duration. I found that perceptual discrimination in whisper, while lower than that for normal speech, was significantly above chance. Together, the perceptual and the production data suggest that whispered speech removes neither the acoustic nor the articulatory distinction between phonologically voiced and voiceless segments. Whisper is therefore not a useful tool for probing the question of articulatory versus acoustic motor control targets.

In the second study, I look at the multiple parameters contributing to relative prominence, to see whether they are controlled in a qualitatively similar way to the parameters observed in bite block studies to contribute to labial closure or vowel height. I vary prominence by eliciting nuclear accents with a contrastive and a non-contrastive reading. Prominence in this manipulation is found to be signalled by f_0 peak, accented syllable duration, and peak amplitude, but not by vowel de-centralization or spectral tilt. I manipulate the contribution of f_0 in two ways. The first is by eliciting the contrastive and non-contrastive readings in questions rather than statements. This reduces the f_0 difference between the two readings. The second is by eliciting the contrastive and non-contrastive readings in whispered speech, thus removing the acoustic f_0 information entirely. In the first manipulation, I find that the contributions of both duration and amplitude to signalling contrast are reduced in parallel with the f_0 contribution. This is a qualitatively different behaviour from all other motor control studies; generally, when one variable is manipulated, others either act to compensate or do not react at all. It would seem, then, that this prosodic variable is controlled in a different manner from other speech motor targets that have been examined. In the whisper manipulation, I find no response in duration or amplitude to the manipulation of f_0. This result suggests that, like in the endoscopy study, perhaps whisper is not an effective means of perturbing laryngeal articulations.

Acknowledgements

> Of all the communities available to us, there is not one I would want to devote myself to, except for the society of the true searchers, which has very few living members at any time.
>
> –Albert Einstein

This work would never have happened without the support and encouragement of Deena, my companion in all things. I hope I can support you as much in the pursuit of your dreams.

Second, though she played very little active part, I must acknowledge the inspiration and motivation that our daughter Kaia gave in the last year of my PhD. Whatever you do with your life, Kaia, I hope you get as much joy from it as I get from my work.

My supervisor, Alice Turk, helped immensely in channelling and encouraging my ideas into productive lines. Your patient guidance, encouragement, and correction has been invaluable to me in honing my skills as a researcher. I could not have asked for a more helpful balance between the freedom you gave me to explore my own ideas and the standards of academic excellence to which you held my every proposition and action.

Bob Ladd, my second supervisor, provided crucial perspective and guidance, particularly on the prosodic theory and literature behind the contrastive emphasis study.

Rob Clark and Jim Scobbie, my examiners, have helped me to refine this work through their critiques.

Bert Remijsen's advice and Praat measurement scripts, as well as his critique of my spectral tilt measurement techniques, improved the acoustic analysis portion of this work immeasurably.

I also owe thanks to Simon King, Cassie Mayo, Sasha Calhoun, Peter Bell, and many other students and staff at the University of Edinburgh, for their input, their expertise, and their willingness to listen as I developed my ideas through presentations and discussions.

The endoscopy study would not have happened without the support and participation of John Esling at the University of Victoria, his students, and the welcoming people at STR-SpeechTech Ltd who allowed us to use their space and equipment for our recordings.

I must thank Michael Dobrovolsky, whose example first inspired me to become a phonetician and whose voice still shows up in my head with encouragement and critique. I will always think of a phonetics lab as a place of real and metaphorical music thanks to you.

The University of Calgary, where I first studied linguistics, remains an important intellectual base for me. I also owe thanks to the Department of Linguistics, and to Steve Winters (head of the phonetics lab) for giving me space while I was writing up this dissertation.

Many other researchers across the world have obliged me with responses to queries. My particular thanks go to John Catford, Louis Goldstein, Sean Fulop, and Harald Baayen for personally responding to my queries about their work.

I thank my parents and my parents-in-law, both for their encouragement and for material support while I worked on this project.

Finally, I could not have done this work without the time given by all of the participants in the experiments and pilot studies that went into this dissertation.

Contents

Declaration ii

Abstract iii

Acknowledgements v

Chapter 1 Overview 1

1.1 Context 1

1.2 Perturbation and compensation 2

1.2.1 Instrumental perturbations 6

1.2.2 Bite block 6

1.2.3 Oral prostheses 8

1.2.4 Non-instrumental perturbations 9

1.2.5 Walking 10

1.2.6 Co-articulatory variation 10

1.2.7 Whisper 11

1.3 Existing work 12

1.4 Empirical questions 13

1.4.1 Phonological voicing in whispered speech 13

1.4.2 Parameter cooperation in signalling contrast 14

1.5 Methodological contributions 14

Chapter 2 Endoscopy measures 16

2.1 Anatomy 16

2.2 Non-video-based measures of glottal opening 19

2.3 Existing measures of video data 20

2.3.1 What has changed? 23

2.4 Correcting image distortions 24

2.4.1 Rectangular pixels 24

2.4.2 Wide-angle distortion 26

2.4.3 Wideangle correction tool 27

2.5 Measurement 31

2.5.1 Glottal aperture 31

2.5.2 Larynx height 32

2.5.3 Obstruction of the vocal folds 35

Chapter 3 Endoscopy study 39

3.1 Opening 39

3.2 Introducing the data 40

3.3 Terminology 43

3.4 Design 45

3.5 Method 46

3.5.1 Equipment 46

3.5.2 Speakers 46

3.5.3 Dataset 46

3.5.4 Recordings 47

3.5.5 Segmentation of the acoustic signal 49

3.5.6 Identifying frames to measure 49

3.5.7 Extraction of corresponding video frame 54

3.5.8 Measurement 57

3.5.9 Annotation reliability test 57

3.5.10 Measurement—summary 59

3.5.11 Statistics used 60

3.6 Results 61

3.6.1 Controls and confounds 62

3.6.2 Distributions and log scaling 62

3.6.3 Cuneiform tubercle width as a covariate 65

3.6.4 Adjustment for cuneiform tubercle obstruction 68

3.6.5 Summary 69

3.6.6 Consonants 70

3.6.7 Normal speech 71

3.6.8 Whispered speech 75

3.6.9 Comparing normal to whispered speech 76

3.7 Perceptual test 80

3.7.1 Dataset 80

3.7.2 Presentation 81

3.7.3 Participants 82

3.7.4 Results and analysis 83

3.8 Discussion 85

Chapter 4 Prosodic measures 87

4.1 Duration 87

4.2 Fundamental frequency 90

4.3 Amplitude 96

4.4 Vowel quality 99

4.5 Spectral tilt 101

4.5.1 Energy bands 102

4.5.2 Regression lines 102

4.5.3 Voicing-based measures 103

4.5.4 Acoustic models 103

4.5.5 Summary of existing spectral tilt measures 108

4.5.6 Modified optimal measure 108

4.5.7 Validation 110

Chapter 5 Contrastive accent study 115

5.1 Experimental question 115

5.2 Speech material 116

5.3 Participants 118

5.4 Recordings 119

5.5 Measurements 120
5.6 Results 121
5.6.1 Statistics used 122
5.6.2 Basic effects 123
5.6.3 Main effect on f_0 123
5.6.4 Main effect on duration 124
5.6.5 Main effect on amplitude 125
5.6.6 No effect on vowel quality 126
5.6.7 No effect on spectral tilt 127
5.6.8 Responses to the whisper manipulation 127
5.6.9 Duration 128
5.6.10 Amplitude 129
5.6.11 Responses to the question manipulation 130
5.6.12 Fundamental frequency 130
5.6.13 Duration 131
5.6.14 Amplitude 133
5.7 Summary 134

Chapter 6 Discussion and conclusions 136
6.1 Voicing contrasts in whispered speech 136
6.1.1 Articulatory study 136
6.1.2 Perceptual study 139
6.1.3 Gestural account 141

6.1.4 The phonology of glottal states 143

6.2 Contrastive accent 143

6.2.1 Question manipulation 144

6.2.2 Whisper manipulation 146

6.3 Summary 146

Appendix A Glossary 148

Appendix B Endoscopy study: Sentences read 150

B.1 Voicing contrasts 150

B.2 Pitch gestures 150

Appendix C Prosody study: Sentences read 152

References 157

List of Tables

3.1 Whisper dataset 47

4.1 Contrastive accent study: control durations 88

4.2 Inconstant control words. 88

4.3 Difference of vowel formants from mean 101

5.1 Contrastive accent study: statements 117

5.2 Speakers and block orders for prosody experiment 119

5.3 Effects of contrastive context on formant frequencies 127

B.1 Whisper dataset 150

B.2 Whisper dataset—pitch sentences (target emphasised) 151

B.3 Whisper dataset—pitch sentences (target not emphasised) 151

C.1 Sentences for prosody experiment—“fastened” 152

C.2 Sentences for prosody experiment—“fastest” 152

C.3 Sentences for prosody experiment—“father” 153

C.4 Sentences for prosody experiment—“marshes” 153

C.5 Sentences for prosody experiment—“morning” 153

C.6 Sentences for prosody experiment—“sources” 153

C.7 Sentences for prosody experiment—“surfing” 153

C.8 Sentences for prosody experiment—“sushi” 154

C.9 Sentences for prosody experiment—“thesis” 154

C.10 Sentences for prosody experiment—“thirsty” 154

C.11 Sentences for prosody experiment—statements, non-contrastive . . 155

C.12 Sentences for prosody experiment—statements, contrastive 155

C.13 Sentences for prosody experiment—statements, non-contrastive . . 155

C.14 Sentences for prosody experiment—statements, non-contrastive . . 156

List of Figures

1.1 Hypothesis space in compensation studies (1) 3

1.2 Ambiguous compensation 4

1.3 Hypothesis space in compensation studies (2) 5

1.4 Winning hypothesis in bite block studies 8

2.1 Midsagittal schematic depiction of vocal tract with endoscope in place 17

2.2 Laryngeal anatomy 18

2.3 Variable views of epiglottis 19

2.4 Distance between vocal processes 21

2.5 Glottal measure from Anastaplo & Karnell (1988) 22

2.6 Rectangular pixel correction 25

2.7 Fuzzy edge close-up 25

2.8 Recorded image of reference grid. 26

2.9 Reference image after barrel distortion was corrected. 30

2.10 Marking the vocal folds and glottal aperture 31

2.11 Different camera positions 33

2.12 Endoscopic image with measurement annotations 34

2.13 Unmeasurable token 35

2.14 Vocal fold obscured by cuneiform tubercle 35

2.15 Cuneiform tubercle obstruction (abstract) 36

3.1 Hypotheses regarding glottal aperture in voicing contrasts. 40

3.2 Hypotheses regarding perception of voicing contrasts in whispered speech. 41

3.3 Spectrograms of [t] and [d] in normal speech. 42

3.4 Spectrograms of [f] and [v] in normal speech. 43

3.5 Spectrograms of [t] and [d] in whispered speech. 43

3.6 Spectrograms of [f] and [v] in whispered speech. 44

3.7 Spectrograms of spoken and whispered fricatives 50

3.8 Spectrograms of spoken and whispered tokens 51

3.9 [s] frame sequence 53

3.10 [z] frame sequence 54

3.11 [p] frame sequence 55

3.12 [b] frame sequence 56

3.13 Glottal aperture distribution for vowels 62

3.14 Glottal aperture distribution for vowels—error bars 63

3.15 Cuneiform tubercle width distribution for vowels 63

3.16 Cuneiform tubercle size distribution for vowels—error bars 64

3.17 Log-scaled glottal aperture distribution for vowels 65

3.18 Log-scaled cuneiform tubercle width distribution for vowels 65

3.19 Log-scaled glottal aperture distribution for vowels—error bars . . . 66

3.20 Log-scaled glottal aperture distribution for vowels (zeros omitted)—error bars 66

3.21 Glottal aperture plotted against cuneiform tubercle width for vowels 67

3.22 Video frames of spoken and whispered vowels 68

3.23 Adjusted glottal and cuneiform tubercle width for vowels 69

3.24 Endoscopic images for “Say peer again”—-spoken 71

3.25 Endoscopic images for “Say beer again”—-spoken 71

3.26 Average glottal aperture over utterance for normal speech 72

3.27 Average log-scaled glottal aperture over utterance for normal speech 72

3.28 Average log-scaled glottal aperture over utterance for normal speech—zeros omitted 73

3.29 Glottal aperture plots for spoken “tier” and “deer”—speaker f5 . . 73

3.30 Glottal aperture plots for spoken “tier” and “deer”—speaker f5 . . 74

3.31 Glottal aperture plots for spoken “fear” and “veer”—speaker m2 . 74

3.32 Endoscopic images for “Say peer again”—-whispered 75

3.33 Endoscopic images for “Say beer again”—-whispered 76

3.34 Average glottal aperture over utterance for whispered speech . . . 76

3.35 Average log-scaled glottal aperture over utterance for whispered speech 77

3.36 Glottal aperture, phonological voicing, and phonation mode means with error bars 78

3.37 Log-scaled glottal aperture, phonological voicing, and phonation mode means with error bars 78

3.38 Consonant glottal aperture and cuneiform tubercle width 79

3.39 Voicing perceptual study—sample slide 82

3.40 Perception scores on endoscopic recordings 84

3.41 Perception scores on endoscopic recordings, subdivided by manner 85

3.42 Diagram of endoscopy results. 86

4.1 Segmented "fastest" token, spoken 89

4.2 Segmented "fastest" token, whispered 89

4.3 Segmented "marshes" token, spoken 90

4.4 Segmented "marshes" token, whispered 90

4.5 Example of pulse marks 91

4.6 MOMEL-smoothed f_0 track 92

4.7 Token with marked-up f_0 contour. 93

4.8 Token with diplophonic f_0. 94

4.9 Token with unusable f_0 contour. 95

4.10 Distribution of nuclear peak frequencies 96

4.11 Distribution of prenuclear peak frequencies 96

4.12 Distribution of interpeak minimum frequencies 97

4.13 Distribution of log-scaled nuclear peak frequencies 97

4.14 Distribution of log-scaled prenuclear peak frequencies 98

4.15 Distribution of log-scaled interpeak minimum frequencies 98

4.16 Amplitude measure example—spoken 99

4.17 Amplitude measure example—whispered 99

4.18 Vowel chart by formant values 100

4.19 Spectral model components. 106

4.20 Measuring spectral tilt—spoken. 107

4.21 Measuring spectral tilt—spoken. 109

4.22 Measuring spectral tilt—whispered. 110

4.23 Spectral tilt of voiced and whispered vowels (dB/octave) 111

4.24 Spectral tilt of voiced and whispered vowels (dB) 112

4.25 Spectral tilt of different vowels (dB) 113

4.26 Spectral tilt of different vowels (dB/octave) 114

5.1 Contrastive accent study—sample slides 121

5.2 Main effect of contrastive context on relative f_0 peak 124

5.3 Main effect of contrastive context on control duration 124

5.4 Main effect of contrastive context on accented syllable duration . . 125

5.5 Main effect of contrastive context on unaccented syllable duration . 126

5.6 Main effect of contrastive context on peak amplitude 126

5.7 No main effect of contrastive context on spectral tilt 127

5.8 Interaction of contrastive context and phonation mode on control duration 128

5.9 Interaction of contrastive context and phonation mode on accented syllable duration 129

5.10 Interaction of contrastive context and phonation mode on unaccented syllable duration 129

5.11 Interaction of contrastive context and phonation mode on peak amplitude 130

5.12 Interaction of contrastive context and sentence type on f_0 131

5.13 Interaction of contrastive context and sentence type on accented syllable duration 132

5.14 Interaction of contrastive context and sentence type on unaccented syllable duration 132

5.15 Interaction of contrastive context and sentence type on control duration 133

5.16 Interaction of contrastive context and sentence type on peak amplitude 134

6.1 Reproduction of figure 2.5 137

6.2 Average log-scaled glottal aperture over utterance for normal and whispered speech—zeros omitted 138

6.3 Endoscopy perceptual response scores—summary 140

6.4 Gestural score for “tier” 141

6.5 Gestural score for “deer” 142

6.6 Revised gestural score for “deer” 142

Chapter 1

Overview

1.1 Context

Linguistics is the study of the patterns of language. We wish to determine what causes lie behind the patterns we observe. Sociolinguists examine social patterns and seek social factors as causes. Morphologists look at morphological patterns and the morpho-syntactic factors behind them. Semanticists look at patterns of meaning; historical linguists look at diachronic patterns of language evolution.

In phonetics, we observe patterns in the physical production of speech. We look for psychological, anatomical, and acoustic explanations for those patterns.

A functional (goal-oriented) perspective is useful in the investigation of motor tasks in general. In reaching tasks, for example, the end-point of the hand has proven to be a key predictive parameter in determining the precise path of the hand. This is opposed to, say, muscle-specific targets which specify individual articulator activation potentials.

When we turn to speech production, a sensible starting point would be to determine whether speech is similar to other motor tasks. Can we identify functional goals which reliably predict articulator trajectories? Unfortunately, speech presents some unique difficulties when we try to set up specific empirical tests. Unlike in reaching or other manual skilled action tasks, not all of the potential functional targets of speech are easily identified. Some, like the listener's eventual understanding of an utterance, are difficult to assess objectively. Others, like the positions of laryngeal and oral articulator groups, cannot be easily observed without intrusive instrumentation (which is likely to alter the behaviour being investigated).

The goals of this dissertation are twofold. First, I pose two empirical questions about the motor control of speech, loosely related by the fact that both revolve around the coordination of laryngeal with supralaryngeal articulations. Second, I develop methods to more reliably measure data produced in such studies.

1.2 Perturbation and compensation

Many linguistic researchers have approached the problem of motor control, using a variety of investigative methods. This research generally uses a perturbation/compensation paradigm. Researchers perturb some aspect of speech (say, restricting jaw movement with a bite block apparatus) and look to see what compensation, if any, occurs (say, extra lip movement to achieve labial closure).

"Perturbation studies have allowed for the examination of a phenomenon known as 'compensatory articulation,' i.e., the achievement of a goal or target production involving compensatory movement by the nonperturbed articulators. The results of these studies have suggested that there is some type of invariant goal in the production of a given speech sound and that variable muscle activations serve to reach that goal." (Baum 1988, p 1662) A control variable acts to keep invariant that which it specifies—a property of the acoustic signal, say, or the relative position of two articulators. By introducing perturbations, researchers can then observe the compensation to determine which properties of the speech are kept invariant. There are two layers of questions to ask within this compensation/perturbation paradigm. First, we ask whether a manipulation of parameter A triggers compensation in parameter B. Second, if compensation is observed, we as what this tells us about motor control.

Figure 1.1 schematizes the hypothesis space used to ask the first question.

A compensation study generally involves two variables that contribute to a linguistic distinction. Experimenters manipulate one variable (eg, jaw opening) leaving the other one (such as lip movement) free to vary. The response of the dependent variable to the manipulation is then observed. There are three types of behaviour that the dependent variable might exhibit (hypotheses a, b, and c in

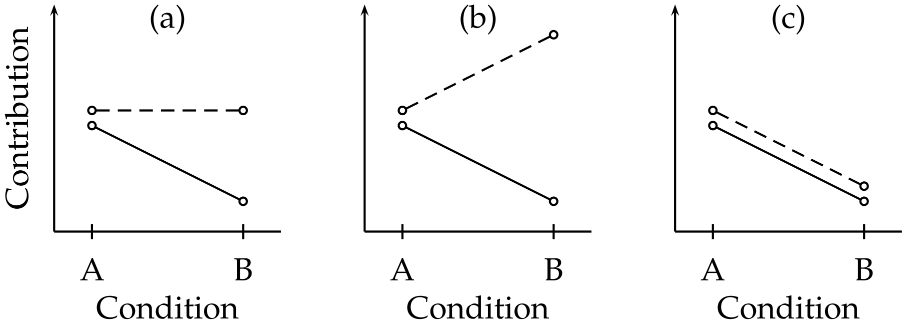


Figure 1.1: Illustration of the contribution of the independent (solid line) and dependent (dashed lines) variables to a linguistic signal. The dependent variable can respond to the manipulation of the independent variable in three ways (discussed in the text).

figure 1.1): it can fail to react at all (a), it can react to the manipulation in the opposite direction as the independent variable does (b), or it can react in the same direction as the independent variable (c).

If there is no correlation between the variables—hypothesis (a) in figure 1.1—this suggests that they are governed by separate motor control variables. Statistically, this is the null hypothesis: we ask whether it is reasonable to rule this possibility out given a particular set of data. The point of motor control variables is to recruit any articulatory resources available to maintain an invariant goal. When a manipulation of one contribution does not trigger a compensatory reaction in the other, the reasonable inference is that these two variables do not cooperate in achieving the hypothesized motor control target.

The next alternative is that there is a significant negative correlation between the dependent variable and the independent variable—hypothesis (b) in figure 1.1. This is exactly the behaviour predicted if both parameters are controlled by the same motor control variable, acting to achieve a single invariant goal. If both variables can contribute to a target, and that target is the object of a motor control variable, then a reduction of one variable's contribution will tend to trigger a compensatory increase in the other variable's contribution—a negative correlation—in order to achieve the target.

Under existing characterizations of control variables, these two hypotheses are the only likely ones: either two parameters are governed by the same control variable, and so exhibit compensation—hypothesis (b)—or they are not governed by the same control variable, and thus don't respond to each other's perturbations—hypothesis (a). The third possibility—that two parameters show a sympathetic rather than a compensatory relationship—would be difficult to explain from within the current motor control. A result consistent with hypothesis (c) would suggest a unique mode of motor control, different from what has been observed in past studies.[1]

We are not only interested in whether there is some sort of compensation—whether, for example, the tongue compensates for jaw immobility in the production of vowels. We ultimately want to know what is driving that compensation. What functional target is suggested by the compensation we observe? In the tongue-jaw example (drawn from the bite-block literature, discussed in section 1.2.2 below), it is tempting to infer that the target is articulatory because the parameter being maintained by the compensation (vocal tract area function) is articulatory. But in this case, the acoustic consequence (the formant structure of the vowel) is also being maintained. Either of these things—the articulatory configuration or the acoustic output—could be invoked as a possible target to explain the compensatory behaviour observed. Figure 1.2 illustrates this situation, drawing on the hypothesis space shown in Figure 1.1.

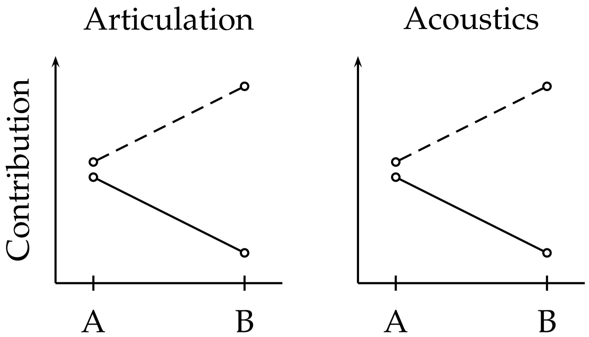


Figure 1.2: Illustration of results that do not help decide between articulatory and acoustic hypotheses (based on schematization from Figure 1.1).

What we need is a manipulation in which the different hypothesized targets predict different behaviours: where a speaker driven by an articulatory target would respond differently to one driven by an acoustic target. Figure 1.3 presents ...

It should be remembered, however, that the useful results schematized in figure 1.3 are not the only possible results in an experiment. Also possible (still omitting the difficult-to-interpret outcome (c) from figure 1.1) are results where both articulation and acoustics show compensation, and where neither show compensation.

[1] On a procedural level, such a result may also point to a poor design, in which the dependent and independent variables are not, in principle, capable of being controlled separately.

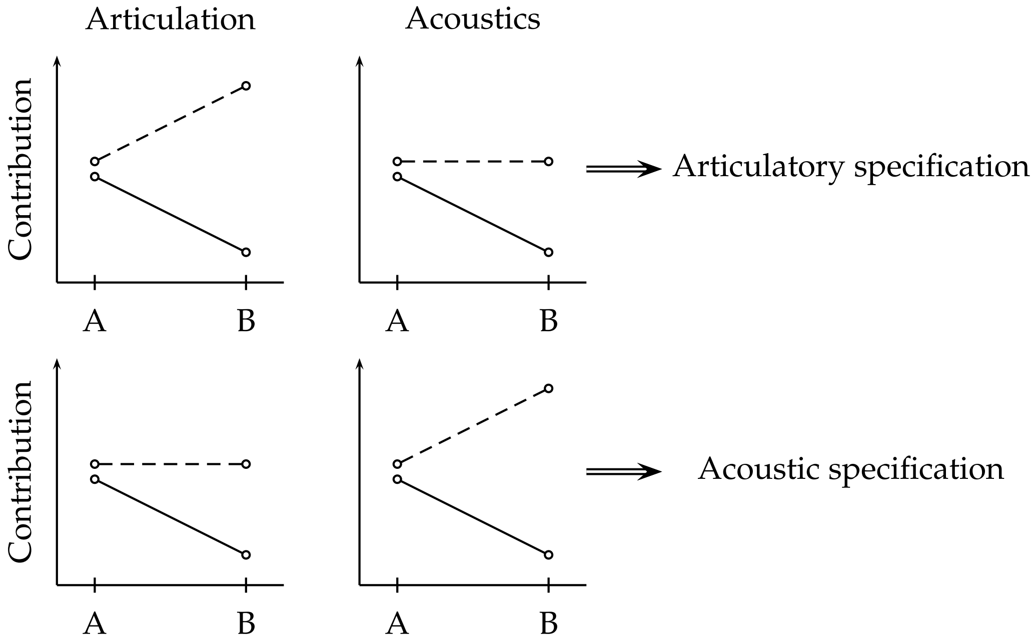


Figure 1.3: Illustration of how to test between different types of control variable specification—results that help decide between articulatory and acoustic hypotheses.

By definition, compensation studies elicit speech in unusual conditions—some form of perturbation is always used. It is prudent to ask whether the patterns thus revealed can be used to make statements about normal speech production.

In a study of Swedish vowels produced with bite-block, Lindblom, Lubker & Gay (1979) say the following:

> *Context-sensitive coding* and *plasticity* of motor control are properties of motor systems in general. To explain the existence of compensatory articulation we can therefore propound the following hypothesis: speakers do so well on the compensatory vowel production tasks because normal speech production programming is indeed "compensatory" (context-sensitive and predictive) in nature. The differences between compensatory and normal articulation do not reside in the choice of different encoding strategies, but rather have to do with extreme versus non-extreme articulatory parameter values. (p159)

Kelso & Tuller (1983) advocate a similar approach:

> Immediate adjustment . . . is a predictable outcome of a dynamical system in which muscles function cooperatively as a single unit. If the operation of certain variables is fixed, as in the bite block, or unexpectedly altered, as in on-line perturbation, linked variables will automatically assume values appropriate to the constraint relation, as long as biomechanical limitations are not violated. In short, dynamical systems—of which speech is an instance—always operate in a mode that one can describe as "compensatory." (p222)

Skilled actions are controlled in such a way that the same motor control strategy can produce successful behaviours in normal and in abnormal conditions. This is confirmed by the predictive success of mathematical models that incorporate such assumptions—such as those for balance in cats (Lockhart & Ting 2007), hand position in human reaching tasks (Flash & Hogan 1985). It is also supported by several of the speech studies mentioned below. In particular, note the agreement between bite block studies (eg Gay, Lindblom & Lubker 1981, Kelso & Tuller 1983) and the walking-and-talking study of Shiller, Ostry, Gribble & Laboissière (2001), described below. The former involve instrumental perturbation; the latter involves simply speaking while walking—hardly an abnormal condition. Both the naturalistic and the instrumental perturbations yield the same compensatory behaviour.

Following is a quick overview of methods used to probe the nature of speech motor control.

1.2.1 *Instrumental perturbations*

In laboratory speech studies, it is unsurprising that many researchers seek to perturb speech instrumentally. They introduce apparatus to either restrict the motion of the articulators or change their shape. The two main types of instrumental perturbations used are bite blocks, which restrict the mobility of the jaw, and oral prostheses, which change the configuration of the vocal tract.

1.2.2 *Bite block*

Bite block experiments perturb jaw movement by blocking the biting motion. Bite blocks can come in two forms. One is a static bite block, which consists of a solid object held between the teeth, usually the lateral posterior teeth, to reduce obstruction of the acoustic speech signal (Lindblom et al. 1979, Lubker 1979, Fowler & Turvey 1981, Gay et al. 1981, Kelso & Tuller 1983, Oller & MacNeilage 1983, Smith & McLean-Muse 1987, Baum, Kim & Katz 1997, Baum 1988, Edwards 1992). The other is a dynamic bite block, in which an apparatus is arranged so that the jaw moves freely at first, but resistance can be applied without warning to the jaw, inhibiting its movement (Folkins & Abbs 1975).

Jaw movement participates in a wide variety of speech articulations—any articulation involving raising of the tongue or lower lip. Interfering with jaw movement therefore affects the production of a large number of speech sounds.

Two important findings come out of bite-block studies with respect to the nature of motor control variables.

The first is that segments are consistently produced with appropriate constrictions in spite of the bite block. Labial consonants exhibit adequate closure (Folkins & Abbs 1975, Smith & McLean-Muse 1987)—the lips move farther to compensate for the reduced jaw mobility. Vowels tend to show normal formant structures immediately on insertion of the bite block (Lindblom et al. 1979, Fowler & Turvey 1981, Gay et al. 1981, Kelso & Tuller 1983). Tracings of the x-ray data collected by Gay et al. (1981) showed that the tongue and lip articulations were adjusted to maintain as normal a vocal tract area function as possible, especially at the points of greatest constriction (the points most relevant to producing natural-sounding segments). Other studies show small but consistent effects of bite block on formant values (McFarland & Baum 1995, Baum et al. 1997, Baum 1999). However, these effects are smaller than would be expected if no compensation occurred, suggesting that *some* compensation is present even when it is not complete. Also, Flege, Fletcher & Homiedan (1988) found that immediate compensation was not complete in the production of [s] and [t] in bite block, in both articulatory and perceptual measures.

The second key finding of bite block studies is that these compensations were immediate, showing little or no delay as might be expected if speakers were replanning their articulatory strategies (Lubker 1979, Fowler & Turvey 1981, Kelso & Tuller 1983).

The compensatory behaviours evident in these bite-block studies suggest a single control variable that specifies labial closure (for labial consonants) or tongue height (for vowels), rather than muscle-specific control variables separately governing the contributions of the jaw and lips.

In terms of the hypotheses schematized in figure 1.1 above, these studies all point clearly to hypothesis (b): compensation is present. The reduced contribution of jaw movement triggers an increased contribution of lip or tongue movement. Note that this hypothesis does not require complete compensation—tongue height, for instance, does not have to be the same relative to the palate for bite-block as for non-bite-block vowels. Schematically from figure 1.4, $j + k$ does not have to equal $l + m$. It is enough that there is *some* negative correlation.

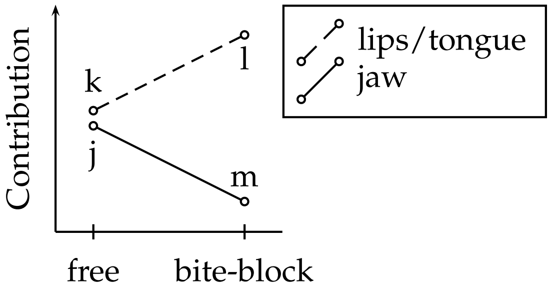


Figure 1.4: Winning hypothesis in bite block studies.

From these studies, we cannot establish whether the true target specified by the control variables is articulatory or acoustic; just that it is composite across multiple related articulators.

1.2.3 Oral prostheses

Another means of instrumentally perturbing speech is by altering the configuration of the articulators. The most common form this takes is the use of an artificial palate such as an electropalatograph (Hamlet & Stone 1976, 1978, Baum & McFarland 1997, McAuliffe, Lin, Robb & Murdoch 2008). By changing the effective shape of the palate, it shifts the point of contact or approximation for consonantal strictures, and it alters the resonating properties of the oral cavity. It also blocks the touch receptors in the palatal skin, thus depriving the speech production mechanism of one potential source of feedback about articulatory performance. Another technique involves using a dental prosthesis to extend the teeth (Jones & Munhall 2003).

Artificial palates used to examine the production of both consonants and vowels caused changes in both articulation and the acoustic signal (Hamlet & Stone

1976, 1978, Baum & McFarland 1997, McAuliffe et al. 2008). Not only this, but the changes were not consistent across speakers. This suggests that there is no coherent compensation strategy being employed. Only after accustomization do productions begin to approximate natural-sounding speech. The time to accustomization varies across studies, speakers, and segment types. Hamlet & Stone (1976) report that some speakers were able to produce normal vowel formants after a week with the prosthesis, while some were not. Studying the production of alveolar consonants with artificial palates, Hamlet & Stone (1978) found that after two weeks speakers' productions were approaching normal (initial overshoot patterns had either disappeared or been dealt with by speakers shifting the place of articulation). Baum & McFarland (1997) accelerated adaptation by having participants read [s]-laden passages over the course of an hour. Both perceptual and acoustic measures suggested that near-normal-sounding [s] was produced by the end of the hour. McAuliffe et al. (2008) observe varied adaptation patterns across their three participants over the 3 hours of their study.

A related study by Jones & Munhall (2003) used dental prostheses that extended the upper front teeth by 5–6 mm. Their goal was "to examine the contribution of auditory feedback to learning a novel acoustic-motor relationship by modifying the vocal tract in a way that did not hinder movement or reduce somatosensory information." (p 533) Unlike artificial palates, dental extension prostheses don't interfere with feedback on contact or pressure. Results for the production of /s/ showed acoustic distortions, which improved over time. (Articulatory data were not recorded.) This agrees with the results of the artificial palate studies.

Speakers do not immediately compensate for the distortion imposed by a prosthesis (as they do with bite-block). Short-term reactions to the presence of the prosthesis are varied and unpredictable—neither the articulatory plan nor the acoustic product are kept invariant. The use of oral prostheses does not, therefore, allow us to examine how speech motor control variables are represented or organized. These studies provide information on the longer-term adaptation by which normal speech is eventually restored, but such longitudinal questions are beyond the scope of the current work.

1.2.4 Non-instrumental perturbations

All laboratory procedures for examining speech carry the risk of generating "unnatural" behaviour—of producing patterns that do not, in fact, represent normal speech production. Are speakers consciously adjusting their targets or otherwise diverging from their normal speech production behaviours? The introduction of experimental apparatus such as bite blocks and oral prostheses compounds this risk. Some researchers have therefore opted to use more naturalistic perturbations to examine compensatory behaviours.

One approach is to simply wait for one parameter (the independent variable) to vary spontaneously, perhaps due to co-articulatory effects, and then observe whether another parameter (the dependent variable) responds systematically to that variation. The walking study (section 1.2.5) and the two co-articulatory variation studies (section 1.2.6) take this approach. The other approach is to directly elicit speech under specific linguistic conditions differing in a relevant (independent) variable, and to look for compensatory effects in another (dependent) variable. This is the approach taken with whisper studies (section 1.2.7).

1.2.5 Walking

Shiller et al. (2001) observed subjects speaking while walking. The varying direction and magnitude of head acceleration (up and down) applies varying loads to the jaw: it wobbles up and down a little relative to the head. This is analogous to the load produced by a dynamic bite block apparatus (Folkins & Abbs 1975, section 1.2.7). Shiller et al. found that, when participants were speaking and walking (rather than just walking) the wobble of the jaw due to acceleration disappeared. Participants were compensating for the acceleration due to walking in order to achieve normal jaw apertures for speech. This study provides valuable validation of the results of bite-block studies (which show compensation to experimentally-induced confounds of jaw movement) as representative of “normal speech”—supporting the claim of Lindblom et al. (1979) and Kelso & Tuller (1983) above that normal speech motor control is inherently “compensatory”, and no special mode of control needs to be invoked to explain speech under perturbed conditions.

1.2.6 Co-articulatory variation

Some studies go one step further even than Shiller et al. (2001), looking only at natural, unperturbed speech and counting on contextual effects within the speech itself to constrain some aspect of production more in some situations and less in others. Perkell et al. (2000, p 246) define the term “motor equivalence” as referring to “the observation that, in multiple tries, the same goal is reached in more than one way.”

Perkell et al. (1993), investigated the production of /u/, one of whose distinctive characteristics is a low F2 frequency. They used an electromagnetic mid-sagittal articulograph (EMMA) to examine the movements of the tongue. Their participants were four male speakers of American English specifically selected from a pool of twice that many on the basis of their prominent use of lip rounding in /u/ production. They elicited /u/ in contexts where co-articulatory effects could be expected to lower the tongue body, raising F2 frequency. They then observed lip rounding, which also affects F2. They found that the tongue-height differences induced by co-articulation were negatively correlated with lip-rounding differences apparently designed specifically to maintain an /u/-like second formant. That is, their participants were using different vocal tract shapes to achieve an invariant auditory goal, depending on the context. This suggests that a motor control variable was acting on an acoustic rather than an articulatory specification, as it was the acoustic and not the articulatory properties of speech which were being kept constant.

A similar study was conducted by Guenther et al. (1999) to investigate the maintenance of the characteristically low F3 of /ɹ/. They elicited it in bisyllabic nonsense words, following different consonants (/b d ɡ v/) and intervocalically. Using acoustic modelling, they identified three acoustic motor equivalence strategies that could be used to maintain the low F3. “Analysis of acoustic and articulatory variabilities revealed that these tradeoffs act to reduce acoustic variability, thus allowing relatively large contextual variations in vocal tract shape for /r/ without seriously degrading the primary acoustic cue.”

If replications with greater numbers of speakers bear out the patterns observed in these two studies, then we will have good evidence for acoustic specifications in the production of the formants studied. Identifying and testing other, similar motor-equivalence strategies could lead to a conclusion that formants of vowels and sonorants are governed by control variables with acoustic specifications.

1.2.7 *Whisper*

Another way to look for compensatory behaviour without intrusive experimental apparatus is to have speakers voluntarily remove a channel of information.

Many studies have looked at whisper—in which vocal fold vibration is absent—to see what (if anything) speakers do to recover the information normally carried in vocal fold vibration. This includes f_0 contours expressing lexical tone (Gao 2002), pitch accent (Nicholson & Teig 2003), intonational information, or sung tune (Meyer-Eppler 1957, Thomas 1969), as well as the phonological voicing contrasts normally signalled (in part) by the presence and timing of devoicing and voicing (Kallail & Emanuel 1984*a*,*b*, Munro 1990, Mills 2003, Higashikawa, Green, Moore & Minifie 2003).

With any methodological tool, such as whisper, it is useful to know exactly which variables are being manipulated and which are not. For example, we know that voice onset time is, strictly speaking, absent from whispered speech. But are all of the glottal distinctions between voiced and voiceless consonants absent in whispered speech? Perkins, Rudas, Johnson & Bell (1976) assume they are when interpreting their finding that shows stutterers are far more fluent in whispered speech. They claim that the reduction in stuttering is because there are fewer gestures to coordinate. On the other hand, several phoneticians have claimed (also without direct empirical evidence) that, in whispered as in normal speech, the glottis is wider for phonologically voiceless segments such as /s/ than for phonologically voiced segments like /z/ (Sweet 1877, 1906, Pike 1943, Malmberg 1963, Abercrombie 1967, Catford 1964, 1977, Laver 1994).

Several studies have looked at the perception of voicing contrasts in whispered speech, finding better-than-chance discrimination (Dannenbring 1980, Munro 1990, Higashikawa & Minifie 1999, Stevens & Wickesberg 2002, Nicholson & Teig 2003, Mills 2003). However, none have eliminated the possibility that non-laryngeal cues, such as duration, are responsible.

1.3 Existing work

The experimental work described in the foregoing sections is far from resolving the basic questions pertaining to speech motor control.

The bite block studies confirm that the control of jaw and lips (in the case of labial consonants) or jaw and tongue (in the case of vowels) is combined. We may infer that this combined control is based on functional targets such as labial closure or tongue height, but we cannot determine from the evidence at hand whether those targets have articulatory or acoustic specifications. Either would generate the patterns observed. The walking study, though a valuable validation of bite-block results, is unlikely as a methodology to tell us anything beyond what bite block studies have already demonstrated.

The research on oral prostheses demonstrates that the motor control system is unable to generate online compensation for alterations of the vocal tract—specifically of passive articulators (palate and upper teeth). In such cases, a process of re-learning is required in order to achieve "normal" productions.

Acoustic studies of motor-equivalent co-articulatory variation provide preliminary support for acoustic goals for vowel and sonorant formants (though much work remains to be done to establish this conclusion).

1.4 Empirical questions

One question which is frequently brought up in speech motor control research is whether the functional targets of speech are specified in terms of articulation or in terms of acoustics.

1.4.1 Phonological voicing in whispered speech

An empirical question was raised in section 1.2.7: is the glottal distinction between voiced and voiceless consonants absent in whispered speech? On the surface, the answer is obviously "yes". After all, phonetic voicing participates in this contrast, either by its presence or absence on the obstruent in question, or by the timing of its onset relative to consonantal release (voice onset time, or VOT).

However, the question is not whether phonetic voicing is used to distinguish obstruents in whispered speech, but whether the glottal distinction that gives rise to voicing contrasts is present. It is on this question that researchers have differed, and it is this question which the current study seeks to answer. By definition, whispered speech lacks phonetic voicing, and thus does not have acoustic contrasts. However, there is a great deal of variability possible in laryngeal articulator configurations even within the constraints of whispered speech. The glottis can be wider or narrower; the vocal folds can be elongated or shortened; transglottal airflow can be greater or less. All three of these parameters could participate in phonetic voicing in normal speech. The current study looks specifically at whether glottal aperture distinguishes phonologically voiced and voiceless obstruents in whispered speech; and if so, how does the distinction compare to that in normal speech.

1.4.2 Parameter cooperation in signalling contrast

Research into compensation has so far looked primarily at segmental speech production: consonants and vowels. The production of suprasegmental elements such as stress, accent, and intonation, is less well-studied. This may be largely due to the fact that the segmental patterns of speech are more well-explored than the suprasegmental patterns.

However, perceptual and acoustic studies have established some patterns of prosody to the extent that they are susceptible to study within the perturbation/compensation paradigm. In this work, I examine the various acoustic cues—particularly f_0 and duration—which are used to signal (and to recognize) contrastive focus on an accented word.

The key question is this: will these prosodic parameters interact in the same compensatory pattern that we have seen, for example, in bite-block perturbations of segmental parameters? Or will the fact that f_0 and duration are generated by different articulatory systems (one laryngeal, the other supralaryngeal) mean that they are coordinated in a qualitatively different way to how jaw and lip movement are coordinated in bite-block speech? The answer to this question will help us understand just how widely the bite block type of compensation is informative about speech production in general.

1.5 Methodological contributions

One theme of the current work, emerging both in the study of voicing in whisper and in the study of parameters signalling contrast, is a focus on articulatory events that occur in the larynx. Whispered speech and phonological voicing are primarily defined by the laryngeal posture employed in the production of the segments in question. Contrastive emphasis involves at least two parameters—f_0 and amplitude—which are primarily controlled in the larynx. (The third primary acoustic parameter implicated in signalling contrast is duration, which is mainly due to the timing of supralaryngeal articulatory events.)

The study of laryngeal motor control is hampered by the inaccessibility of laryngeal structures to direct instrumental observation. Recordings with intrusive instruments such as endoscopes or electrodes used in electromyography are expensive to acquire and uncomfortable for participants, necessitating short recording sessions. Most data from which laryngeal events are inferred is indirect—primarily acoustic data. Inferences of laryngeal features such as glottal vibration characteristics (via spectral tilt) or pharyngeal/epiglottal retraction (via formant structures) depends on models that map articulations to acoustics. The validity of inferences can never exceed the validity of the acoustic models—all of which employ some simplifying assumptions.

A secondary goal of this work is to improve upon existing measurement techniques to get precise, reliable measurements of parameters related to laryngeal motor control. Chapter 2 introduces the data gathered in the endoscopy study, and presents a new technique (drawing on existing methods from the literature) for acquiring well-controlled glottal aperture values from endoscopic video data. Chapter 4 outlines the acoustic measures taken for the study of contrastive emphasis. In addition to outlining measures of duration, f_0, and amplitude, this chapter presents a thorough review of the literature on spectral tilt. Measures of spectral tilt have been used for many things, but often tend to be used as a proxy for certain properties of the glottal waveform. I identify the measure best-suited to the current study (in which the source waveform is of particular interest), and introduce a slight modification designed to expand its applicability.

CHAPTER 2

Endoscopy measures

While fiberoptic endoscopy has been available for phonetic research for several decades—it was first introduced to the phonetic research community by Sawashima & Hirose (1968)—a technique for obtaining reliable, controlled quantitative measurements from endoscopic recordings remains elusive.

This chapter gives the reader an overview of the issue of measuring endoscopic data. A brief review of anatomy (section 2.1) is followed by a survey of measures described in existing research (section 2.3). Next, I present the techniques used to correct distortions in the images (sec:endoscopy-correcting-distortions), followed by the procedure for extracting measurements (sec:endo-measurement).

Chapter 3 presents the details of data acquisition, and the results of measuring that data with the technique described here.

2.1 Anatomy

The basic vocal anatomy and the positioning of the lens are common to all endoscopic studies.

Endoscopic recordings give a view of the larynx from above. The position of the endoscope relative to the larynx is illustrated in figure 2.1.

Figure 2.2 illustrates the anatomical structures of the larynx that are typically visible in endoscopic images. Endoscopic images are oriented with the posterior of the larynx approximately at the top. All marked structures aside from the epiglottis and the cuneiform tubercles are composed of soft tissue. The epiglottis

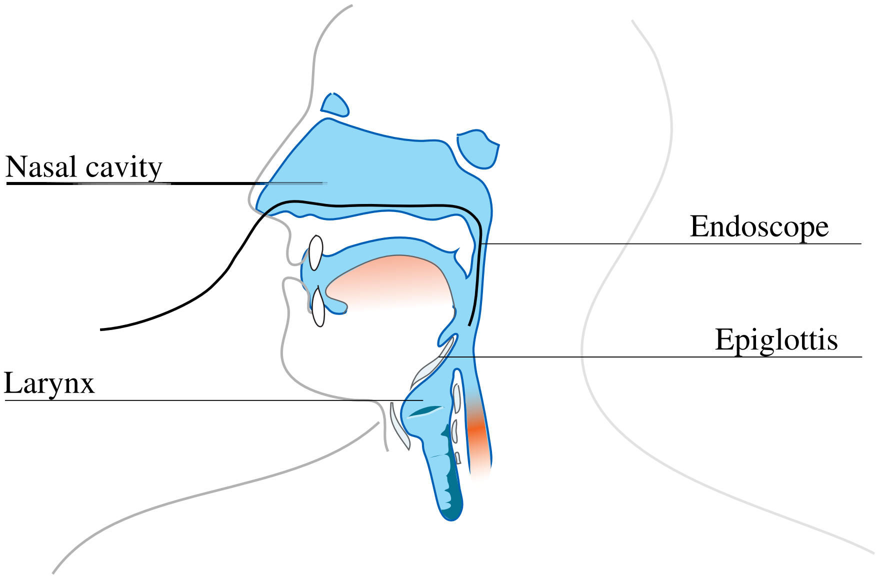


Figure 2.1: Midsagittal depiction of vocal tract with endoscope in place, adapted from image in Wikimedia Commons "http://commons.wikimedia.org/wiki/Image:Particulate_danger-it.svg" (public domain image), posted by user "Xander89".

and cuneiform tubercles are cartilaginous structures covered by skin. Note that, although circles are used to identify the cuneiform tubercles in the figure, they are rarely circular in appearance. In this figure, they appear roughly triangular; in figure 2.12 later in this chapter (a different speaker), they appear as long ovals. The arytenoids—another pair of skin-covered cartilages—are not clearly visible in this type of data. They lie inferior to the cuneiform tubercles (from this perspective, they are beyond the cuneiform tubercles).

The crescent-shaped upper edge of the epiglottis is sometimes visible, sometimes not, as shown in figure 2.3. Its base is always visible—the bump at the anterior end of the vocal folds is the epiglottal tubercle at the base of the epiglottis. The posterior portions of the aryepiglottal folds are always visible, though their points of connection to the epiglottis are often not visible (as in figure 2.2). The other labelled structures—cuneiform tubercles, ventricular folds (false vocal folds), and vocal folds—are usually visible. (The ventricular folds can obscure the vocal folds during particularly constricted productions. Also, see Esling (2002), Esling & Harris (2003), and Edmondson & Esling (2006) for examples of endoscopic images in which the epiglottis is so retracted as to completely obscure the glottis—such articulations are beyond the scope of the current work.)

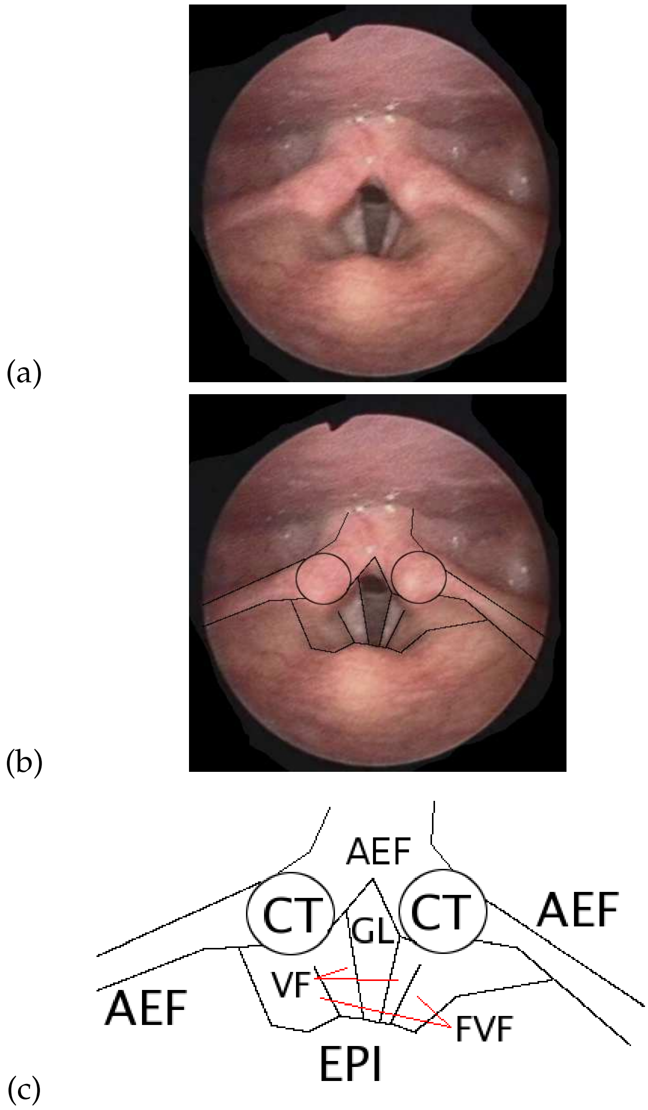
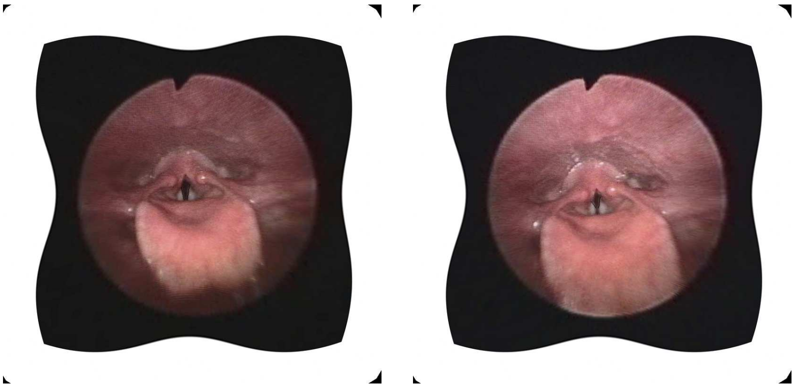


Figure 2.2: Illustration of laryngeal anatomy (a) as seen through an endoscope, (b) with key structures outlined, and (c) labelled abstraction. Label key: GL = glottal opening, VF = vocal folds, FVF = false vocal folds (ventricular folds), EPI = epiglottis, AEF = aryepiglottal folds, CT = cuneiform tubercles.



(a) (b)

Figure 2.3: Frames with the upper edge of the epiglottis visible (a) and not visible (b).

The structures labelled *cuneiform tubercles* in the present work are so identified following Edmondson & Esling (2006). However, Benguerel, Hirose, Sawashima & Ushijima (1978) identify the same structures as the *corniculate cartilages*. In his discussion of laryngeal anatomy, Zemlin (1988) describes the corniculate cartilages as resting on the superior apexes of the arytenoid cartilages (p106), while the cuneiform cartilages are embedded in the aryepiglottic folds (p108), somewhat lateral to the corniculate cartilages (Zemlin's figure 3-13, p109). Unfortunately, he does not provide a labelled image of the endoscopic view to decisively distinguish the two structures in an endoscopic image. It is therefore possible that these structures are mislabelled in the current work. I retain Edmondson & Esling's attribution, as the labelled structures in figure 2.2 seem to agree more with Zemlin's description of the cuneiform cartilages than his description of the corniculate cartilages. Note that the property relevant for the current work—namely, that they are of constant size and can thus serve as a normalizing reference for distance (section 2.5 below) holds regardless of the name we use.

2.2 Non-video-based measures of glottal opening

There are methods of observing (or inferring) the magnitude of glottal opening which do not involve gathering video data.

Transglottal illumination (also known as photoglottography) uses a similar setup to endoscopy, with the addition of a light sensor applied externally to the front of the throat below the larynx. (ie Lisker, Abramson, Cooper & Schvey 1969, Baer, Lofquist & McGarr 1983, Gracco & Löfqvist 1994, Hess & Ludwigs 2000) This detects the amount of light transmitted through the larynx from the endoscopic tube. (This can be done in conjunction with enodscopic recordings.) This has the advantage of generating a single time-varying signal. That signal correlates with the size of the glottal opening (the wider it is, the more light passes through). However, being a one-dimensional signal, it provides no scope for dealing with confounding variables such as relative height of the larynx and supraglottal obstruction by the cuneiform tubercles, ventricular folds, or epiglottis (see below for discussion of these). An additional confound, not experienced in the current endoscopic study, is reported by Gracco & Löfqvist (1994). They had to discard 14% of the data from one of their 3 participants because the participant's tongue kept obstructing the light source.

A laryngograph (or electroglottograph) uses electrodes attached on either side of the larynx to obtain a trace of glottal contact area over time. (ie Baer et al. 1983, Henrich, d'Alessandro, Doval & Catellengo 2004) This is far less invasive, as the sensors are external to the vocal tract. The electrodes measure resistance across the larynx, which varies inversely with the amount of contact between the vocal folds. More contact means less resistance. This measure does not suffer from the problem of obstruction as video and transillumination techniques do.

Measures taken using these different techniques are reported to be highly correlated with each other. Because of this, I decided to us only one: endoscopic video recordings. I felt that this technique offers the richest raw data from which to extract meaningful measures while adjusting for important confounding factors.

2.3 Existing measures of video data

While some important qualitative work on endoscopic data has been done and continues to be done without numerical measurements (Esling 2002, Esling & Harris 2003, Edmondson & Esling 2006), the ability to quantify observations for statistical analysis is necessary for the testing of more specific or subtle hypotheses.

Various instrumental setups can be used in phonetic research to generate endoscopic images.[1] For example, high-speed analysis of individual vocal fold vibrations (thousands of frames per second) (Hayden & Koike 1972, Tanabe, Kitajima,

[1] I will not cover studies that use rigid oral endoscopes, as they dramatically interfere with natural speech articulation. However, many of the points made about the measurement of data from fiberoptic nasal endoscopes would apply equally to data from rigid oral endoscopes.

Gould & Lambiase 1975) differs from normal-speed recordings (25-30 frames per second) used to observe variations in laryngeal configuration across segments (Kagaya 1974, Hirose, Lee & Ushijima 1974, Hirose & Ushijima 1978, Benguerel et al. 1978, Benguerel & Bhatia 1980). Different again is the use of stroboscopy with normal-speed recordings to capture some high-speed characteristics of vocal fold vibration without needing a high-frame-rate recorder (see, for example Anastaplo & Karnell 1988).

Most researchers measure the distance between the vocal folds, taken at a natural landmark—the vocal processes of the arytenoid cartilages (Hirose et al. 1974, Kagaya 1974, Benguerel et al. 1978, Hirose & Ushijima 1978, Iwata, Sawashima, Hirose & Niimi 1979, Sawashima & Park 1979, Benguerel & Bhatia 1980). This is relatively posterior, and therefore tends to be the point of greatest aperture along the length of the vocal folds. This is illustrated in figure 2.4, which depicts a [p] at the moment of release, spoken by a female participant in the current study.

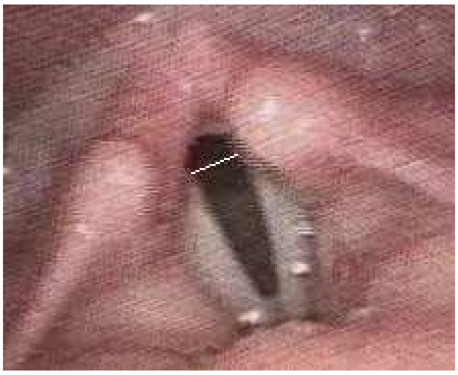


Figure 2.4: Common glottal aperture measure (see text) comprising apparent distance between the vocal processes of the arytenoid cartilages.

While less time-consuming than any of the alternatives, this measure has an important disadvantage. It is vulnerable to two key intra-recording confounds—varying camera-to-larynx distance, and cuneiform tubercle obstruction of the posterior portion of the vocal fold. This does not preclude it being a useful measure, but a measure that deals with these confounds would be more powerful—particularly for studying vocalizations in which larynx height and vocal fold obstruction vary systematically with variables being manipulated (as they do in whispered speech).

Other studies have reported a variety of alternative measures of glottal opening.

In a procedure designed for analysis of high-speed endoscopic video, Hayden & Koike (1972) use a series of five points along the length of each vocal fold to calculate both maximum glottal aperture and total glottal area. A similar measure is presented by Tanabe et al. (1975), who measure glottal width at six points along the length of the vocal folds. Tanabe et al. (1975) point out one advantage of making measurements across frames of a single glottal cycle: "The vertical movement of the larynx does not effect the measurement, since the phonation is a vowel phonated at a constant pitch, and the time span covered on the film is short." (p80)

Anastaplo & Karnell (1988) present a measure which is designed specifically to validate EGG measures. From images acquired using stroboscopy, they identified three points along the midsagittal line of the glottis (depicted in figure 2.5, adapted from their figure 4, p1885):

> Points P and A represent the *posterior* and *anterior* borders of the glottis during maximum opening. Point C marks the posterior-most point of *contact* along the superior surface of the vocal folds during periods of glottal closure. Relative length of glottal opening (G) was measured as $G = PC/PA$.

In other words, their measure G was the portion of the antero-posterior length of the glottis which was open. Because it is a ratio, it is already effectively normalized for camera-to-larynx distance.

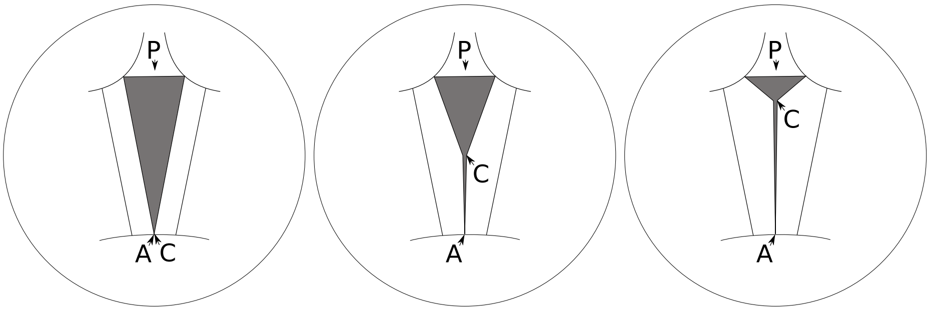


Figure 2.5: Glottal measure from Anastaplo & Karnell (1988), from their figure 4 (p1885). Points indicated (P, C, and A) are described in the text. The leftmost diagram indicates a maximum G measure of 1.0; the two to the right represent decreasing values (about 0.5 and 0.2).

Benguerel et al. (1978) identify a further difficulty with the distance between the vocal processes. They note that this distance is generally a very small distance on the film, so it is particularly vulnerable to small errors in measurement. To mitigate this error, they add a second measure—the distance between the corniculate cartilages (what I identify as the cuneiform tubercles in the current work—see earlier comment on labelling these structures).

A recent study ignores distance measures altogether. Dailey, Kobler, Hillman, Tangrom, Thananart, Mauri & Zeitels (2005) opt instead to measure glottal angle—the angle at which the anterior ends of the vocal folds intersect. This measure has the advantage of being independent of camera-larynx distance, and so not requiring normalization. Unfortunately, this measure cannot cope with glottal configurations where the vocal folds are closed along their anterior portion, then widen in a posterior glottal chink (such as in the middle diagram in figure 2.5. In such cases, glottal angle increases as glottal opening *decreases* (ie, as the glottal chink becomes shorter)—exactly the reverse of the normal relationship between glottal angle and glottal opening. The measure is thus unsuitable for high-speed analysis of vocal fold vibrations (see, for example, the measure of Anastaplo & Karnell 1988 mentioned above) or for analysis involving whispered or breathy states of the glottis.

2.3.1 What has changed?

The reader may now wonder, with this wealth of well-used and accepted measurement techniques already available, why one would wish to develop a whole new procedure for quantifying glottal aperture in endoscopic images. After all, the data itself is largely the same.

The main change is one of resources. As noted by previous researchers (Hayden & Koike 1972, Tanabe et al. 1975), endoscopic measurements are time-consuming. When most of the above-cited research was performed, only the most rudimentary aspects of the analysis could be aided by computers. Now, however, the data is recorded and stored digitally, and significant parts of the processing can be fully or partially automated using image analysis and editing software. Thus, for the same amount of research time and with substantially less capital costs for video and computer equipment and software, we can obtain much better-quality data than previous researchers could.

For example, because of its small size, a fiberoptic endoscope will always need a wide-angle lens and will thus distort the image to some degree (cf Dailey et al. 2005). It is now possible to digitally correct this distortion using a combinationof computer-aided measurement and fully-automated processing scripts (see section 2.4.3).

Building on the existing strategies, I have developed a procedure for removing distortions and extracting measurements which accounts for much more of the confounding variation and thus gives us much cleaner data than previous techniques yielded. This should improve the power and reliability of the statistical inferences that we can draw from endoscopic data.

2.4 Correcting image distortions

Two types of image distortion must be removed before the frames can be measured. The first, rectangular pixels, relates to the encoding format of the video data. The second, barrel distortion, is a consequence of the lens optics. They are removed separately, as described below.

2.4.1 Rectangular pixels

Video images recorded in NTSC format[2] have "rectangular" pixels—their width is not equal to their height. This has its roots in the pre-digital foundations of NTSC video-coding, where vertical distance was quantized (images were scanned and displayed as rows) but horizontal distance was not quantized (the horizontal information was encoded in an analog signal, not a digital one. In the still images extracted from the video recordings, the circular field of view through the endoscope is not a circle. The field of view in the uncorrected image displayed on the right of figure 2.6 is about 338 pixels high and 369 pixels wide—about 10% wider than it is high.

Exact pixel dimensions of the field of view are difficult to obtain, due to blurred edges in the digital image. Figure 2.7 illustrates the fuzziness of edges in the image. Depicted is a close-up of the edge of the field of view, marked (as in most images) by at least two rows of pixels of decreasing brightness. It is not clear which level of brightness should be taken as the true edge. For quantifying the current distortion, I took the midpoint of the region in doubt—halfway between the full brightness of the main image and the full darkness of the border—as the point to measure as the edge.

[2] NTSC is a standard video format in North America, where the current data were recorded. Other regions use different standards—such as the PAL format in Europe. They all tend to have rectangular pixels, but with different proportions.

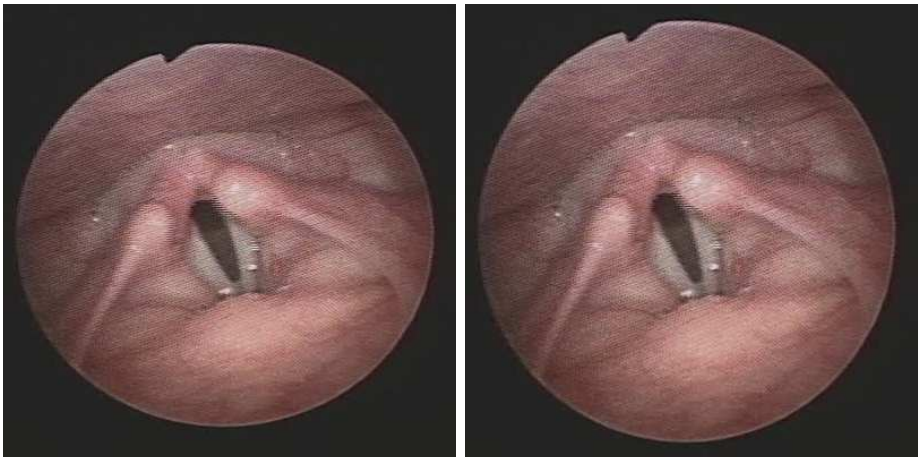


Figure 2.6: Before (left) and after (right) correction for rectangular pixels.

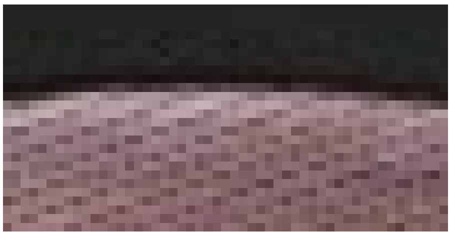


Figure 2.7: Close-up illustrating fuzziness of the edge of the field of view—the boundary between the dark upper portion and the lighter lower portion of the figure.

Measurements on multiple frames converged on a height-to-width ratio of 10:11—the field of view was about 10% wider than it was tall. I was unable to find a clear peer-reviewed technical reference to verify that the pixel aspect ratio of NTSC television images is 11:10. However, various online resources did back up this figure. See for example "http://www.activeservice.co.uk/video/pixels/page2.htm" (viewed 9 June 2009). From these observations, we can infer that a pixel in an NTSC-standard image (and on an NTSC-compliant television) has a height-to-width ratio of 11:10. The extracted still images were processed on a computer; computer monitors use square pixels (height-to-width ratio 1:1). This explains our observation that, when displayed on a computer monitor, each pixel in an NTSC-recorded image (and thus also the picture as a whole) appears short—the height is reduced 10% relative to the width.

Because distance measures made in our analysis rely on linear Cartesian geometry, a transformation was applied which restores the original proportions: the vertical dimension was extended by 10% on all extracted frames using a batch processing script. The image on the right in figure 2.6 shows the result after rectangular pixel correction was applied to the raw image on the left.

2.4.2 Wide-angle distortion

The optical properties of the lens produce barrel distortion (bowing-out) in the image. This interferes with measurements of distance and angle—distances and sizes are non-linearly reduced toward the edge of the field of view. This problem is described by Dailey et al. (2005), from whom the following correction method is derived.

A reference image was generated containing a 15 mm by 15 cm grid composed of 1 cm by 1 cm squares. This reference image was recorded with the experimental apparatus at a distance of 50 millimetres, so that the reference squares covered the entire field of view, as seen in figure 2.8.

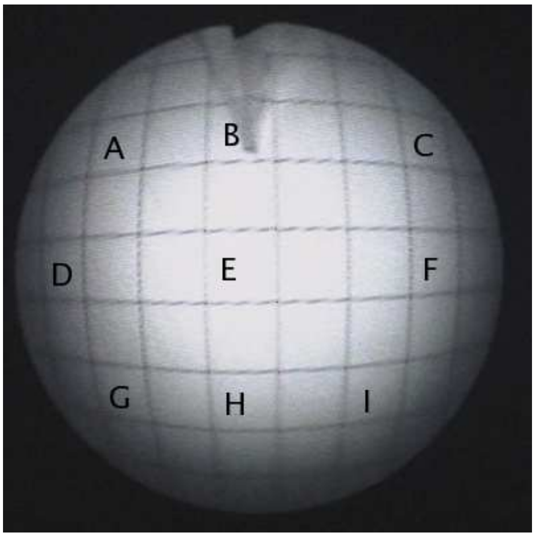


Figure 2.8: Image of reference grid with barrel distortion (image has been corrected for rectangular pixels as described in section 2.4.1). The letters have been added to identify squares used to measure distortion.

The barrel distortion was corrected using the following procedure. All manipulations were performed using the GNU Image Manipulation Program (GIMP 2007) with an extension specifically designed to deal with barrel distortion (Hodson 2007).

In order to quantify the distortion of the original images, measurements were taken to determine how "square" the squares in the image were. Two properties characteristic of squares were measured: corner angles and edge lengths.

Squares have 90° angles at their corners, and all sides have equal length. Under barrel distortion, distances closer to the edge of the image are reduced relative to those in the centre and angles are warped, some becoming more acute and some more obtuse.

Nine squares in the image were selected based on their distribution and the visibility of their corners. The squares were selected from the centre and from the edge of the endoscope's field of view—labelled "A" through "I" in figure 2.8. The coordinates of the corners of each labelled square were recorded.[3]

From these coordinates, corner angles and edge lengths were calculated. In the uncorrected image (figure 2.8), the mean edge length was 44 pixels, with a standard deviation of 6 pixels—14.6% of the mean length. Angles ranged from 71° to 113°, with a standard deviation of 11.3°, which is 12.6% of the average angle of 90°.[4]

With standard deviations of 14.6% in distances and 12.6% in angles, where the ideal is zero variation in both, any measurements taken on the uncorrected data would be greatly confounded by the distortion, especially if the measurements are made in different parts of the field of view (see, for example, figure 2.11 below). Also, the measurement error will mean data will have a large residual variance in statistical tests, compromising the ability to detect real effects.

2.4.3 Wideangle correction tool

This section provides a brief summary of the behaviour of the correction tool used to remove the barrel distortion (Hodson 2007).

The tool constructs a new image in which information from each point in the original image is moved toward or away from the centre according to formulas 2.1 and 2.2.

$$x_{orig} = x_c + k \cdot \left(1 + ar^2 + br^4\right) \cdot (x_{mod} - x_c) \tag{2.1}$$

[3] Coordinates of a point can be easily recorded in Gimp like so: a 1x1 pixel brush is selected for use with the pencil tool, and placed over each corner in turn. The coordinates are read off from the information panel in the lower left corner of the image window.

[4] The mean internal angle was 90°—a geometric certainty when dealing with four-sided shapes of this sort. Mean angle could not therefore be used to quantify distortion.

$$y_{orig} = y_c + k \cdot \left(1 + ar^2 + br^4\right) \cdot (y_{mod} - y_c) \tag{2.2}$$

In these formulas, (x_{orig}, y_{orig}) is a pixel location in the original image and (x_{mod}, y_{mod}) is the location of the corresponding pixel in the modified image.

The coordinates (x_c, y_c) define the centre of the effect (which may not be the centre of the image). The coordinates (x_c, y_c) are derived from parameters "xshift" and "yshift" in Hodson's tool, each of which takes a value from -100 to 100, representing the full width w and height h (respectively) of the image. The conversion from xshift to x_c is $x_c = w * (100 + xshift)/200$ and from yshift to y_c is $y_c = h * (100 + yshift)/200$.

The scaling parameter k was always set to 1.0 in this work.

The value r represents the Euclidean distance between the centre of the effect (x_c, y_c) and the pixel (x_{mod}, y_{mod}), relying on the assumption that pixels are square (they have the same height as width). This assumption is satisfied, as the rectangular pixel correction was applied before wideangle correction (see section 2.4.1).

Finally and most importantly, the parameters a and b specify the magnitudes of the second- and fourth-order distortions (respectively)—the actual wide-angle effect itself. These correspond to the "main" and "edge" parameters of Hodson's tool. The "main" (second degree) correction factor affects the centre of the image more strongly; the "edge" (fourth degree) correction factor affects mainly the edges of the image.

It may seem, intuitively, that the equations should give x_{mod} and y_{mod} (unknowns) in terms of x_{orig} and y_{orig} (knowns). However, equations 2.1 and 2.2 are used to determine where the information for each pixel in the modified image will come from in the original image. For a given pixel in the new image, x_{mod} and y_{mod} are known (they define its location in the new image); it is x_{orig} and y_{orig} (giving the location in the old image where that pixel's information should come from) that need to be calculated.

A minor change in a single calculation parameter was made to the wide-angle tool's source code before compiling it, due to the extreme nature of the barrel distortion introduced by our lens. (Without this correction, the maximum values for the parameters were insufficient to completely remove the barrel distortion from the test image.) I changed the following line in the source code (in the function `wide-angleSetupCalc()`):

```
calcVals.mult_sq = vals.square_a / 200.0;
```

to

```
calcVals.mult_sq = vals.square_a / 100.0;
```

The value `vals.square_a` represents the user-defined parameter "main", which can vary from -100 to +100. The value `calcVals.mult_sq` represents the value a in equations 2.1 and 2.2. This alteration doubles the effect of the "main" (quadratic) correction parameter. A test confirmed that a value of 50.0 with the modified version produced the same effect as a value of 100.0 with the original version. The value of a in equations 2.1 and 2.2 could range from -1.0 to 1.0; the value of b could range from -0.5 to 0.5.

Exploring the parameter settings of the correction tool, I was able to subjectively approximate a good correction (the lines looked approximately straight). To further improve "squareness" before experimental measurements began, the following "slope-climbing" algorithm was followed. It was performed manually, as automated edge detection for determining distortion was not available.

- Each of the two parameters "main" and "edge" in the correction tool (a and b in equations 2.1 and 2.2) was adjusted up and down from the current best-known values by a minimal amount, producing 4 candidate modified images.
- The distortion was measured for each of these four images as described above.
- The candidate with the least variance was taken as the new best-known values, and the procedure was repeated.
- When no more improvement was obtained, the current best result was deemed the best possible.

A final corrected image is shown in figure 2.9. Notice how the correction affects the square border of the original. Because barrel distortion compresses distances more toward the edge of the image, the correction involves stretching distances near the edge proportionally more than those closer to the centre. The optimum correction output by this algorithm increased the overall mean edge length by 25% to 54.9 pixels. This increase is not a problem, as distances are available only in arbitrary units (pixels). It is a necessary side-effect of the correction, which stretches the borders of the image but leaves the centre relatively unchanged. There was a 1.2 pixel standard deviation in lengths—about 2.1% of the mean length. Angles varied from $86°$ to $95°$, with a standard deviation of $1.8°$ (2.0%).

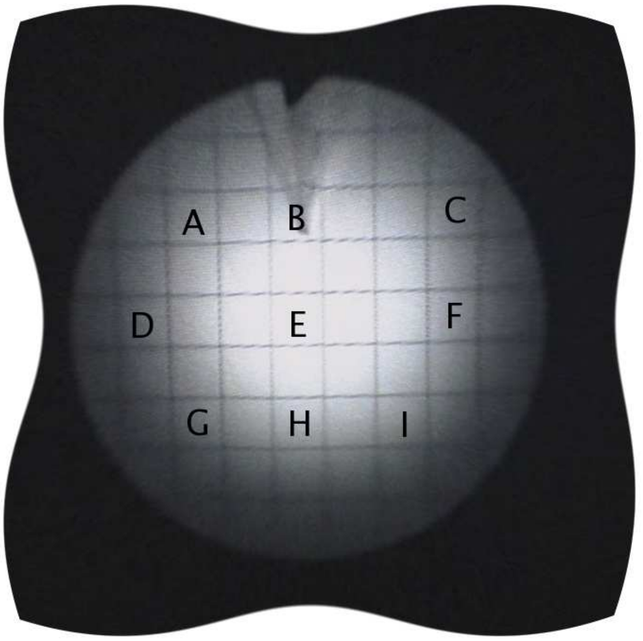


Figure 2.9: Reference image after barrel distortion was corrected.

This correction was obtained with the following parameter settings:

- xshift $= -1.5$ (slight horizontal correction for centre of effect)
- yshift $= 0.0$ (field of view already vertical centred in image)
- main $= -85.0$ (positive values introduce or increase wide-angle distortion; negative values remove or decrease it)
- edge $= 90.0$ (this positive value adjusts for overcorrection at the edges of the field of view from the "main" parameter)

Remember that the parameter "main" was used with a slightly-modified version of the original tool from Hodson (2007).

These measurements show a marked improvement over the original image. The ideal solution would yield no variance in either edge length or corner angles. Unfortunately, these measurements are vulnerable to ambiguity, particularly at the edge where lighting is less optimal and the lines are several pixels wide after correction. It is unlikely that a perfect correction is attainable in practice. The correction obtained—reducing variance in both angle and edge-length to less than 3%—is the best we can achieve with the techniques available to us.

2.5 Measurement

After correction for rectangular pixels and barrel distortion, video frames were measured for glottal aperture.

2.5.1 *Glottal aperture*

For the investigation of voicing, we need a measure to represent glottal aperture. The techniques described here owe much to previous quantitative work on endoscopic images (described in section 2.3 above).

Following Iwata et al. (1979) and Sawashima & Park (1979), I chose to measure the maximum visible distance between the vocal folds—almost invariably achieved at the posterior end. Figure 2.10 illustrates how this is done.

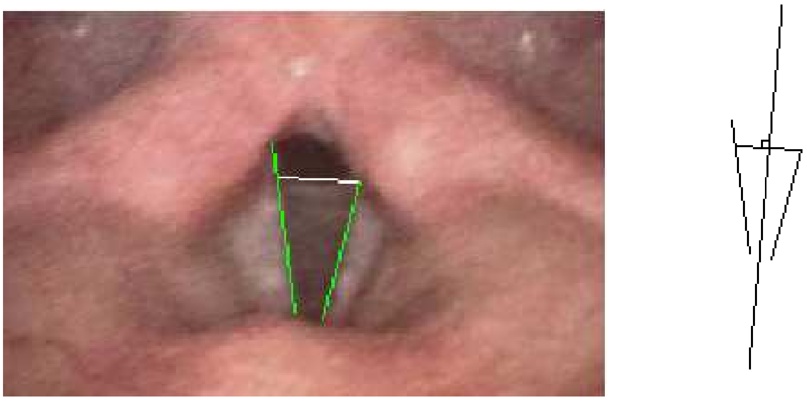


Figure 2.10: Marking the vocal folds and glottal aperture. (Same token as in figure 2.2. The tracing on the right shows hypothetical bisector between lines of vocal folds, from which angle of measurement for glottal aperture is derived (see text).

First the visible portions of the vocal folds were marked: straight-line approximations were drawn between the visible end-points. The angle (in degrees from horizontal) and length (in pixels) of each straight line were recorded.[5] The angle of a hypothetical line bisecting the vocal fold lines was calculated. Glottal aperture was measured at the widest visible point, perpendicular to the hypothetical bisector. For the image in figure 2.10, that distance is 33.1 pixels, marked with an almost-horizontal white line. The tracing to the right of the image shows the hypothetical bisector used to determine the angle at which glottal aperture was measured.

2.5.2 *Larynx height*

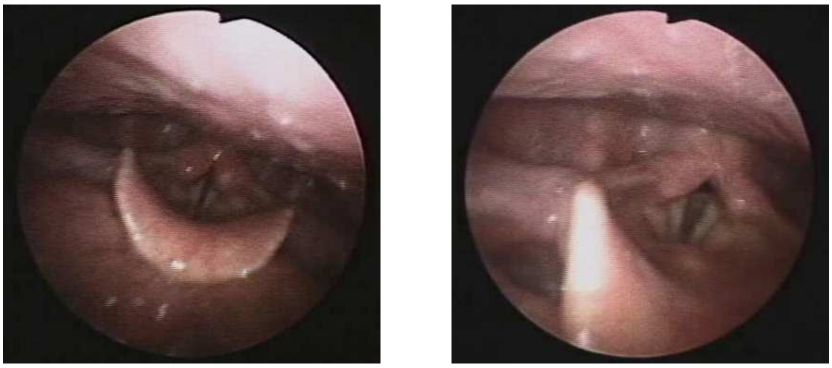
This pixel distance depends not only on the actual glottal aperture, but also on the distance between the camera and the glottis. Significant variation in camera-glottis distance was observed even between nearby tokens produced by a single speaker, as illustrated in figure 2.11. No method was available to calibrate the visual data of varying object-to-camera distance to a fixed referent of known size. Tanabe et al. (1975, p 80) exploit the optical properties of a lens in conjunction with a metric ruler to calibrate sizes based on constant camera-to-vocal-fold distance. Nobody reports adapting this method to recordings with varying camera-to-object distances. Fujimura, Baer & Niimi (1979) describe a system which uses twin fiberscopic lenses—one inserted through each nostril—to construct a dual image from which stereoscopic measurements can be used to determine absolute distances between the lenses and objects in the joint field of view. However, such equipment was not available to us.

Briefly, there are several reasons why larynx height would vary relative to the endoscopic lens within a recording. On the one hand, the lens itself might be moved by the velum, the tongue, or the experimenter (if, for example, the participant becomes uncomfortable and needs things shifted). On the other hand, the larynx itself moves for several reasons. Vocal fold tension is controlled, in part, by forward/backward rotation of the thyroid cartilage on the cricoid cartilage, which involves some vertical movement. Also, laryngeal and epiglottal sphinctering mechanisms are engaged in whisper (Gao 2002, Esling 2002, Esling & Harris 2003); this increased activity seems to involve, either directly or as a common side-effect, raising of the larynx.

Several techniques for measurement of endoscopic images have been outlined in the past (Hayden & Koike 1972, Tanabe et al. 1975, Sawashima 1979, Iwata et al.

[5] The "Measure" tool in GIMP (GIMP 2007) provides the Cartesian distance in pixels between two points and the angle from the horizontal of the line connecting them.

1979); unfortunately, they all rely on the assumption (explicit or implicit) that the larynx does not move vertically with respect to the camera.



Distant, centred

Closer; camera shifted to the left

Figure 2.11: Different positions of the camera relative to the laryngeal structures from the same subject.

A measurement of the cuneiform tubercles in figure 2.11 shows that those in the right-hand image appear about 30% larger than those on the left. The cuneiform tubercles are cartilaginous structures, and do not change size or shape from one utterance to the next; this difference in apparent size can be attributed to the varying distance between them and the camera.

Because the cuneiform tubercles are approximately the same distance from the camera as the vocal folds, we can assume that the effect of camera distance on apparent glottal aperture is proportional to its effect on apparent cuneiform tubercle size. I recorded apparent cuneiform tubercle size on all tokens as a proxy for camera-to-vocal-fold distance. This measure was used in the statistical analysis (see section 3.5.11) to account for possible confounds due to varying larynx height.

The size of each cuneiform tubercle is measured as its maximum visible width. While the cuneiform tubercles illustrated in figure 2.10 are roughly circular, those of other speakers (such as that shown in figure 2.12) are much wider in one dimension than the other. Taking the maximum visible width helps to maintain consistency within a speaker's data. (Note that the wide variations in anatomy between speakers means that glottal aperture cannot be directly compared across speakers, even after incorporating the normalizing measurement of cuneiform tubercle width.)

The tracing of visible boundaries (vocal folds, cuneiform tubercles) is the most subjective step in the measurement procedure, as it requires a human annotator to select discrete boundaries in non-discrete data. Further work using these techniques should employ inter-annotator tests of consistency or, ideally, machine-based edge-detection algorithms wherever possible. For the current investigation, such resources were unavailable. A test of intra-annotator consistency was performed for all measures, and is reported in section 3.5.9.

Figure 2.12 shows a frame of speech with all the measurements marked out. In total, seven values were gathered from each still image: two vocal fold lengths, two vocal fold angles, one glottal aperture, and two cuneiform tubercle widths. A token was discarded if any of these could not be measured. For example, in

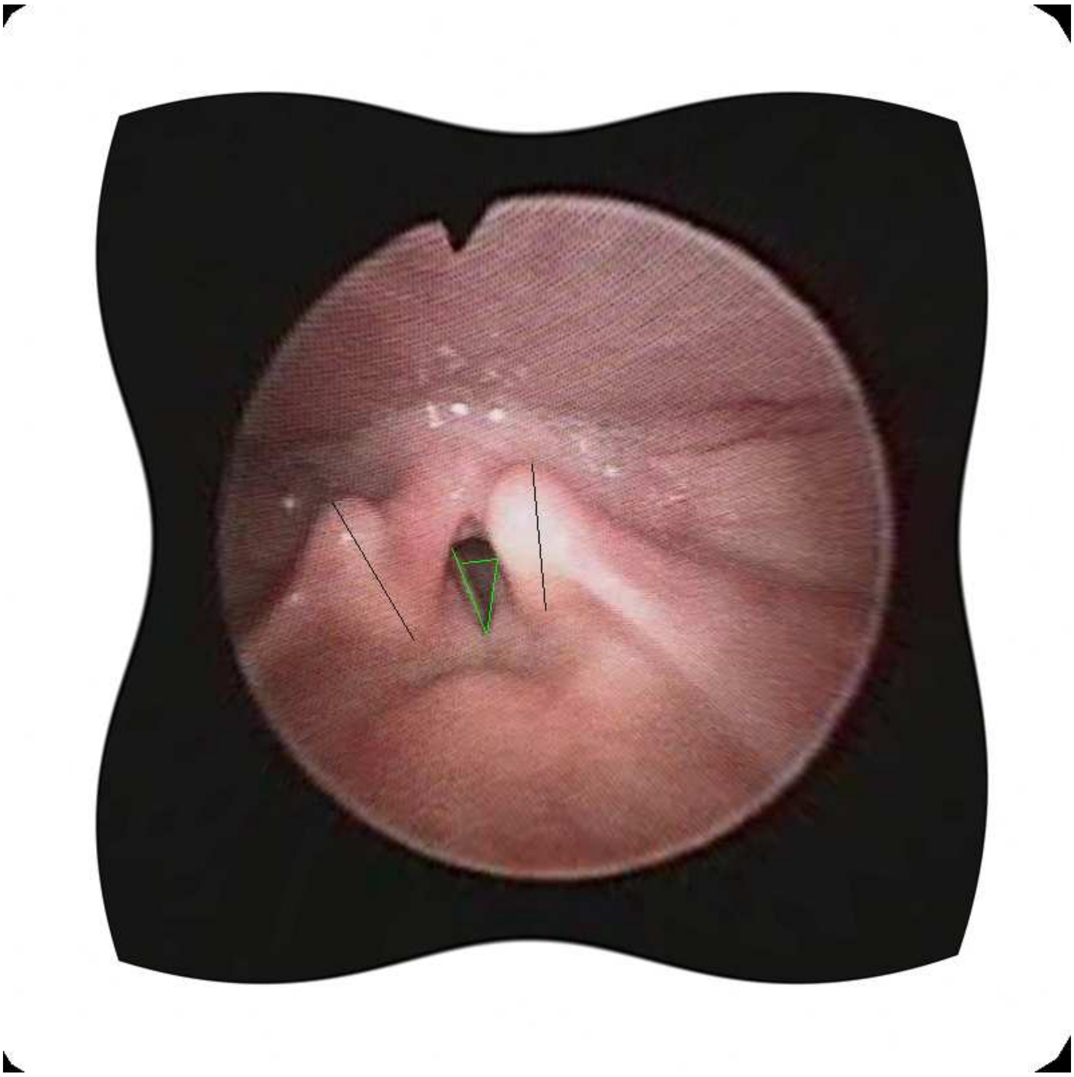


Figure 2.12: Token of whispered [p] with measurement annotations. The long dark lines mark the measurements of cuneiform tubercle size; green lines show the measurements of the vocal folds and the glottal aperture between them.

the token of whispered [f] shown in figure 2.13, the right vocal fold is not visible and so glottal aperture cannot be measured.

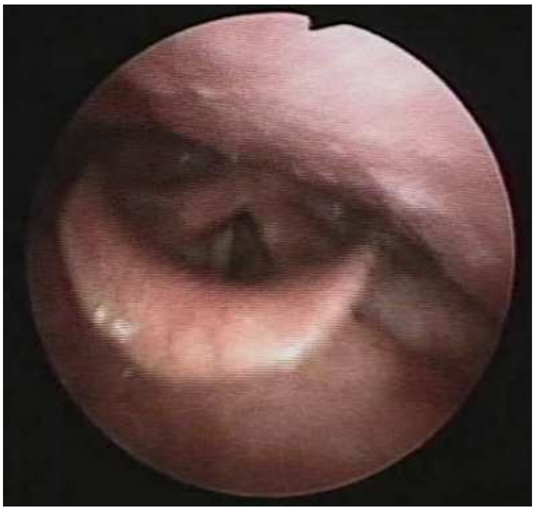


Figure 2.13: Unmeasurable token: right vocal fold is not visible.

2.5.3 *Obstruction of the vocal folds*

One final confounding influence is the tendency, in many tokens, for one or both vocal folds to be partially obscured by a cuneiform tubercle, as illustrated in figure 2.14.

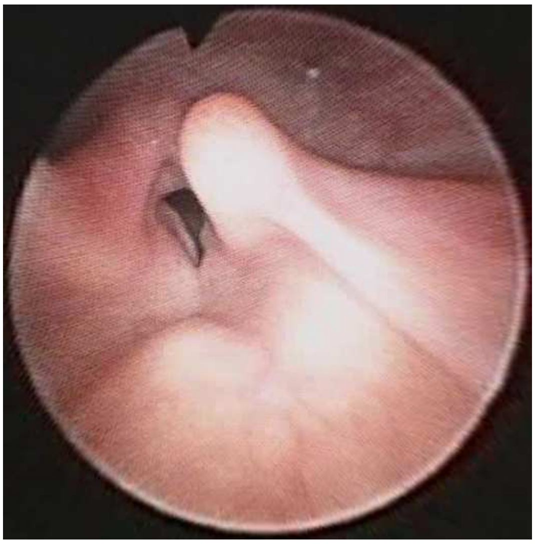


Figure 2.14: Example of vocal fold partially obscured by cuneiform tubercle.

This obstruction is schematised in figure 2.15. The measurements we can make are shown as line segments BC (observed glottal aperture) and as AD and AC (visible vocal fold lengths). The measurement we desire—actual glottal aperture—is shown as line segment DE. Vertex E is the unknown point in these measurements.

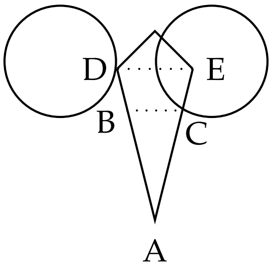


Figure 2.15: Illustration of cuneiform tubercle obstruction of the camera's view of glottal aperture. Vertices are labelled.

This schematization depends on the following reasonable approximations. (1) It assumes that the vocal folds are straight lines. In fact, they are often bowed slightly inward along their length, but they are nearly straight. This assumption allows us to use simple straight-line trigonometric properties in our correction. (2) It assumes that each of the triangles ($\triangle ABC$ and $\triangle ADE$) is an isosceles triangle (the sides corresponding to the vocal folds have equal length). In practice, the vocal folds are not of the same length, but by measuring glottal aperture (the third side) perpendicular to the midline between the other two sides, we ensure that the measured values do form an isosceles triangle. (3) It assumes that triangles $\triangle ABC$ and $\triangle ADE$ are similar (in the mathematical sense that the internal angles are the same, and therefore the sides of one triangle are each larger or smaller than the corresponding sides of the other by the same ratio). This is an automatic consequence of the first two abstractions.

One property of the schematization in figure 2.15 is that the ratio of the actual glottal aperture (DE) to the measured glottal aperture (BC) is equal to the ratio of the long vocal fold measurement (AD) to the short vocal fold measurement (AC). We begin using congruent sides of the similar triangles:

$$\frac{DE}{BC} = \frac{AD}{AB} \tag{2.3}$$

but since the triangles are isosceles, we know that $AB = AC$, so we substitute to get:

$$\frac{DE}{BC} = \frac{AD}{AC} \tag{2.4}$$

which, after multiplying both sides by BC, yields a definition of the actual glottal aperture in terms of the measured values:

$$DE = \frac{AD \cdot BC}{AC} \tag{2.5}$$

It bears restating that this formula is based on a stylized simplification of the actual geometry of the endoscopic images. Also, each multiplication and division of measured values has the effect of compounding the measurement error, so the derived measure will be even less precise than any of the individual raw measurements. On the other hand, by calculating an adjusted glottal aperture measurement using this formula, we may be able to reduce the confounding influence of cuneiform tubercle obstructions of the vocal folds. The adjustment reduces precision (increases variance), but to the extent that the geometric assumptions on which it is based are correct, it improves accuracy—it brings the mean tendency of the measure closer to what we're actually interested in measuring.

The magnitude of the adjustment in different subsets of the data can be determined by looking at the ratio between the adjusted and the plain glottal aperture measures. This ratio is never less than 1.0, because the fraction by which the plain measure is multiplied to yield the adjusted measure has a larger numerator than denominator *by definition*. Over all measurements, the ratio ranges from 1.0 (no change) to 6.0 (the adjustment yields a value six times as great as the plain measure), with a mean of 1.33 (indicating a 33% increase between the plain and adjusted measures). The magnitude of adjustment does not seem to differ greatly between consonants (mean=1.32) and vowels (mean=1.33). A substantial difference is observed between normal and whispered speech. The mean ratio in normal-speech tokens is 1.19; in whispered speech, the mean ratio is 1.45. Clearly, tokens of whispered speech tend to be adjusted by a greater relative amount than tokens of normal speech. This is in line with the subjective impression of the author that there is more cuneiform tubercle obstruction in whispered than in normal speech.

In the absence of an independent, objective verification of the validity of the adjustment, both the plain and the obstruction-adjusted glottal aperture measures are reported, and the results of each are interpreted with the competing considerations of precision and accuracy in mind.

CHAPTER 3

Endoscopy study

3.1 Opening

As mentioned in section 1.4.1, this study seeks to establish the usefulness of whispered speech as a tool for investigating the nature of motor control variables, and to discover the patterns of laryngeal control of glottal aperture in normal and whispered phonological voicing contrasts.[1].

The main empirical question in this study is whether glottal abduction gestures are observed for phonologically voiceless obstruents in whispered speech. An auxiliary question is whether, in the presence of such gestures, voicing minimal pairs are perceptually distinct.

The first question is expressed diagrammatically in figure 3.1. Figure 3.1(a) illustrates what is already known about glottal aperture in normal speech: vowels are produced with a narrow "voiced" glottal posture, voiced obstruents retain the same glottal aperture, and voiceless obstruents exhibit a wide "voiceless" glottal aperture. If glottal abduction gestures are present, as predicted by Sweet (1877, 1906), Pike (1943), Malmberg (1963), Abercrombie (1967), Catford (1964, 1977), and Laver (1994), then we expect a behaviour like 3.1(b). In this situation, phonologically voiced obstruents pattern with the vowels and have a "whispered" glottal aperture, intermediate between voiced and voiceless. Phonologically voiceless obstruents show the wide "voiceless" glottal aperture that they

[1] I am grateful to Professor John Esling of the University of Victoria for his collaboration in running this experiment and for commenting on the design. The supervising physician and the participants were paid from Dr. Esling's SSHRC grant #410-2003-1624. STR-SpeechTech Ltd in Victoria, who housed the laryngoscope equipment, allowed us to use their recording space for the recordings.

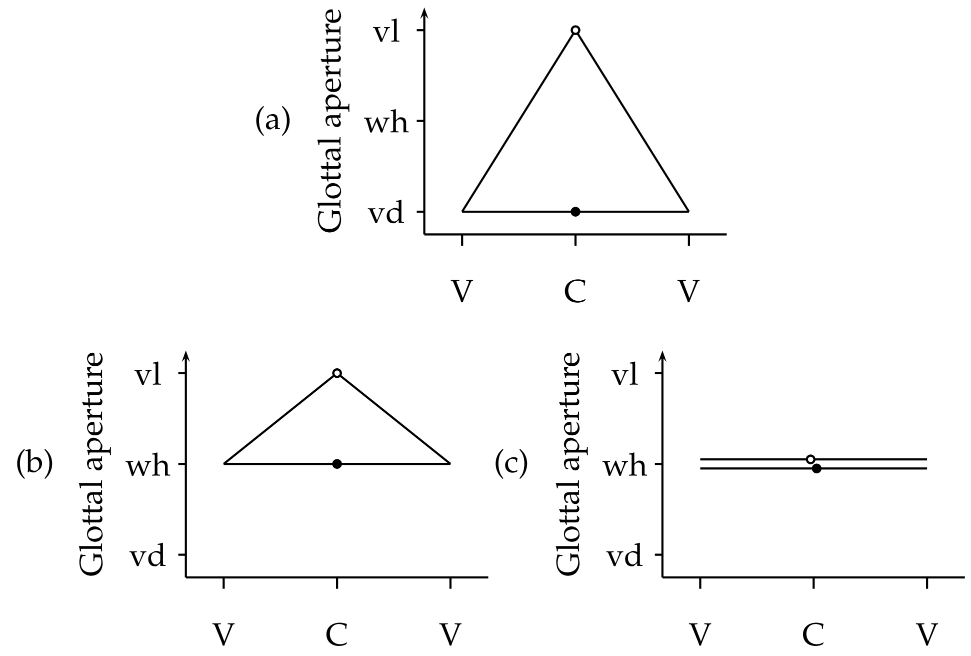


Figure 3.1: Hypotheses regarding glottal aperture in voicing contrasts: (a) normal speech; (b) whispered speech—abduction gesture present; (c) whispered speech—no abduction gesture. Open circles represent phonologically voiceless obstruents; closed circles represent phonologically voiced obstruents.

do in normal speech. If such gestures are absent, as asserted by Perkins et al. (1976), our results will look like 3.1(c): glottal posture will remain "whispered" throughout the utterance.

The perceptual question is really a continuum, the endpoints of which are shown in figure 3.2. On the left, we see perceptual performance completely unhindered by the absence of phonetic voicing. On the right, we see perceptual performance reduced to chance. Two empirical alternatives are open to us: testing whether perception of whispered speech is significantly worse than perception of normal speech, and testing whether perception of whispered speech is significantly better than chance. Previous studies have already established that perception of voicing contrasts in whispered speech is worse than it is in normal speech; we will therefore be asking the latter question.

3.2 Introducing the data

I will begin with a brief visual introduction to the contrasts in question using spectrograms. The spectrograms in figure 3.3 illustrate the difference between [t]

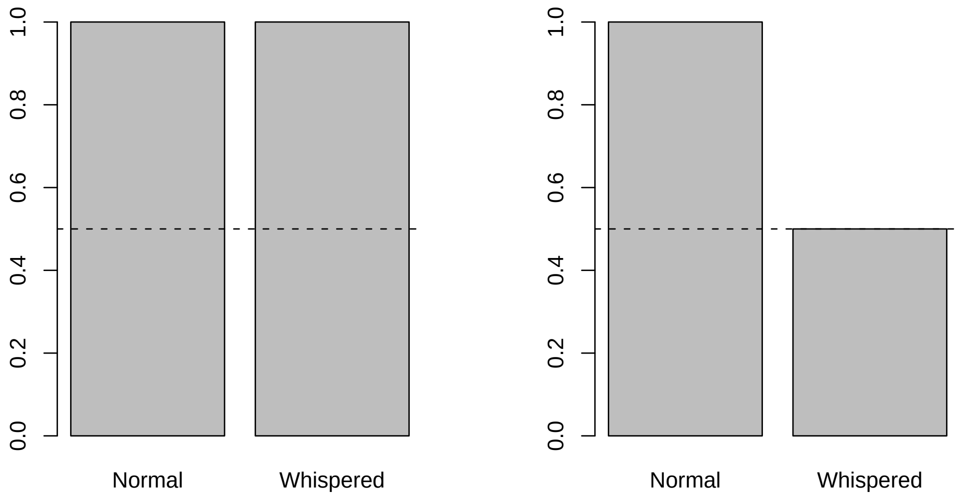


Figure 3.2: Hypotheses regarding perception of voicing contrasts in whispered speech. The vertical axis is proportion of correct responses by listeners in a forced-choice task with two choices. The dashed horizontal line represents chance performance.

and [d] in normal speech. The most striking differences are that the voicing ends earlier for [t] (at about the same time as the formants cut off) and resumes later (lagging significantly behind the consonant release) than for [d]. One can also see that the duration of [t] is somewhat greater than for [d] (reflecting a more general pattern for English: see Crystal & House 1988, Mills 2003).

The spectrograms in figure 3.4 illustrate the difference between [f] and [v] in normal speech. Here, the main difference again revolves around voicing—the [f] has no voicing during its closure, while the [v] is voiced throughout. The same durational difference is present here as in the stops.

Compare the spectrograms of the [t]-[d] distinction in normal speech (figure 3.3) with those of the [t]-[d] distinction in whispered speech (figure 3.5). In the complete absence of voicing, we no longer have aspiration as such. Note, however, that the high-frequency noise post-release in the whispered [t] is comparable to that seen in normal speech, and might be used to acoustically distinguish it from the whispered [d], which lacks this high-frequency noise post-release. The durational difference observed between these two segments in normal speech is also apparent in whispered speech.

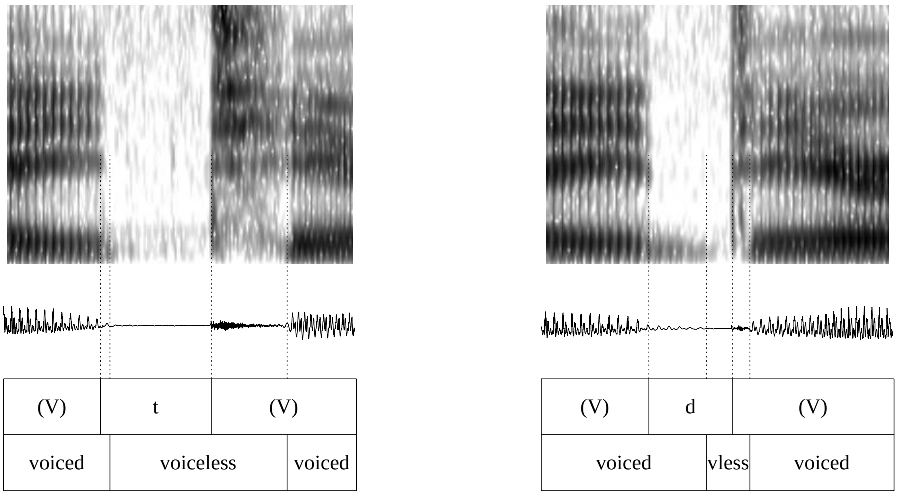


Figure 3.3: Spectrograms and waveforms of [t] and [d] in normal speech. Each item shows 350 ms of speech. The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

Now compare the spectrograms of the [f]-[v] distinction in normal speech (figure 3.4) with those of the [f]-[v] distinction in whispered speech (figure 3.6). Here, the contrast is less obvious. The durational difference is preserved, but very little else leaps to mind to distinguish [f] from [v] in these whispered examples. There may be a slight tendency for [f] to have marginally greater amplitude than [v].

Because the primary acoustic reflex of the glottal aperture differences in normal speech—the presence or absence of voicing—is absent in whispered speech, an acoustically-specified control variable is not expected to preserve the glottal aperture difference in whisper. An articulatorily-specified control variable, on the other hand, is expected to produce the same articulatory gestures in different contexts, regardless of the immediate acoustic consequences.

Therefore, if the glottal aperture difference reflecting phonological voicing contrasts in normal speech is also present in whispered speech, it would suggest an articulatory specification for the control variable governing glottal aperture. If there is no glottal aperture difference, it would suggest an acoustic specification.

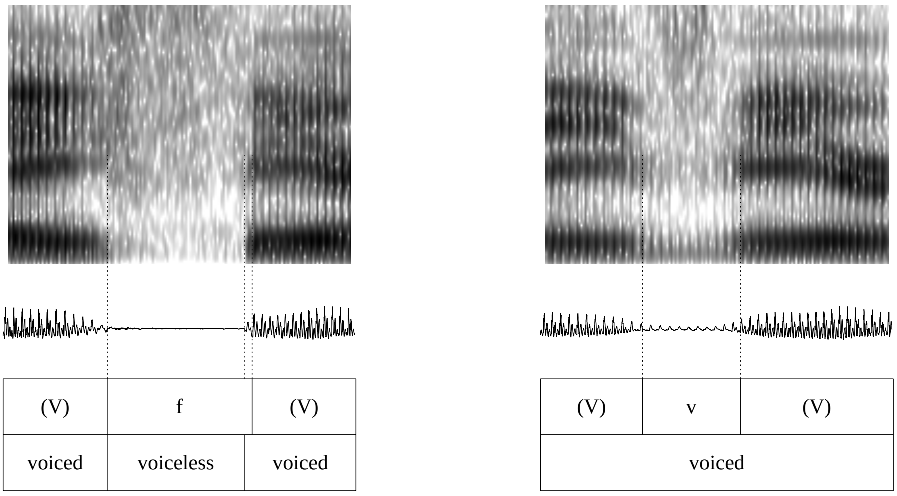


Figure 3.4: Spectrograms and waveforms of [f] and [v] in normal speech. Each item shows 350 ms of speech. The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

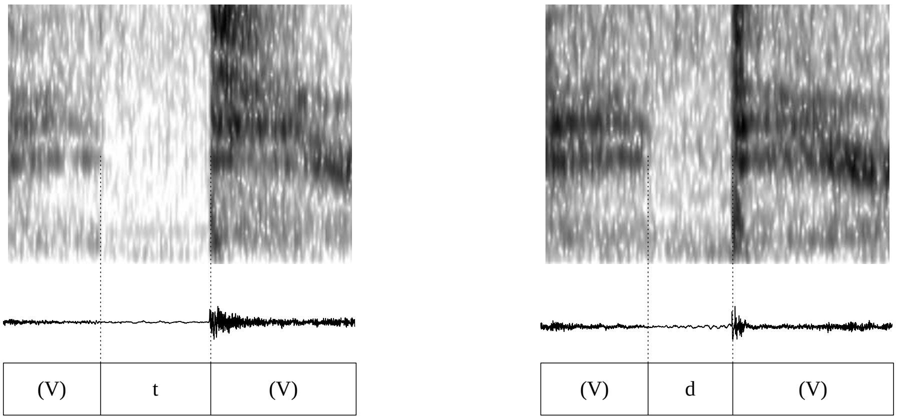


Figure 3.5: Spectrograms and waveforms of [t] and [d] in whispered speech. Each item shows 350 ms of speech. The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

3.3 Terminology

In this study, I use whispered speech as a tool for examining the coordination of glottal articulations and their relations to supraglottal articulations. There are two terminological ambiguities that arise in discussing whispered speech. This section presents the problem (vagueness and polysemy in terms as used in the

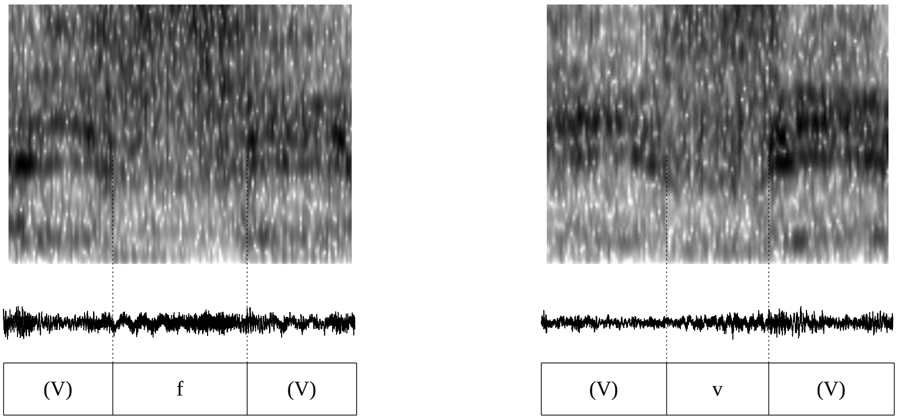


Figure 3.6: Spectrograms and waveforms of [f] and [v] in whispered speech. Each item shows 350 ms of speech. The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

literature) and the particular definitions adopted for this work. A glossary (Appendix A) is provided for reference, defining these and other useful terms.

Sweet (1877) introduces the first ambiguity—in the term “whisper”:

> The popular and the phonetic use of the term ‘whisper’ do not quite agree. Whisper in popular language simply means speech without voice. Phonetically speaking whisper implies not merely absence of voice, but a definite contraction of the glottis. (p5)

Because both senses of the term *whisper* will be used side-by-side in the discussion of this study, we need to establish a clear terminological convention. The two senses of whisper mentioned by Sweet and others (e.g. Abercrombie 1967, p 28) are both, in fact, “phonetic”, in the sense that both describe behaviours relevant to phonetic inquiry. I use *whispered speech* to denote whisper in Sweet’s “popular” sense—as a property of utterances without voice, whose primary mode of vocal tract excitation is turbulence from airflow through a constricted glottis.This is opposed to *normal speech*, by which I mean speech whose excitation is primarily from voicing. I use *whisper phonation* or simply *whisper* to denote the specific glottal state which produces non-voiced, turbulent glottal excitation as a sound source for the vocal tract filter. Normal speech contains segments of voicelessness (produced with voiceless phonation). Crucially, whispered speech may likewise contain segments of non-whispered voiceless phonation on segments like /t/ and /f/. It is this latter possibility that is the object of the current study.

A second terminological ambiguity arises when we put phonological voicing contrasts alongside discussion of phonetic voicing states. This study deals with "voicing" in two different senses: the phonetic sense (a glottal state) and the phonological sense (a feature which represents the contrast between phonemes such as /p/ and /b/). In this dissertation, I use the terms *voiced* and *voiceless* to refer to glottal states, unless explicitly modified as *phonologically voiced* and *phonologically voiceless*.

3.4 Design

This study is designed in two parts. First is the recordings of normal and whispered speech using a nasal endoscope. This part is designed to answer the question of whether there is a glottal aperture difference in whispered speech between phonologically voiced and voiceless consonants, and to compare any difference to that seen in normal speech. The second is the use of the audio recordings from the first part in a perceptual study, to determine the perceptual discriminability of the voicing pairs in normal and whispered speech.

Note that this is a non-instrumental study, in the sense used in section 1.2.4, because the experimental perturbation (whispered speech) is non-instrumental. The presence of an endoscopic tube with a diameter of 4 mm in the nasal passages and the lens just above the epiglottis causes discomfort in participants, and almost certainly has a effect on the speech produced. However, the endoscope was in place for all recordings. The analysis will compare whispered speech utterances with the endoscope to normal speech utterances with the endoscope. Any difference (or lack of difference) observed between the conditions can be attributed to the experimental manipulations, not to the (assumed constant) effect of the endoscope's presence in the vocal tract.

3.5 Method

3.5.1 Equipment

A flexible fiberoptic nasal endoscope (Kay 9100 Rhino-Laryngeal Stroboscope system with Olympus ENF type P3 scope with a 28mm lens and Panasonic GP-US522 camera) was used to capture live video of the larynx while a directional microphone placed approximately twenty centimetres from the speaker's mouth captured the acoustic signal. All data were initially recorded onto a digital tape. The primary recording medium was a mini DV tape using a Sony GV-D1000 NTSC miniDV recorder attached to the endoscope (data encoded in MPEG-1 format: video 720x480 pixels, 30 nominal frames per second; audio 44.1 kHz stereo, 224 kbps). Recordings were later transferred to DVD for storage, using Adobe Premiere software (Adobe 2001).

3.5.2 Speakers

Participants in this study were all native speakers of Standard Canadian English. They were paid CAN$50 (except for the primary researchers, the author and John Esling, who were not paid) for their participation.

Ten speakers participated in the study. Recordings were completed for nine; one was unable to proceed due to acute discomfort during insertion of the endoscope. Of the nine, two were phoneticians familiar with the details of the study (the author and John Esling), one was a linguist unfamiliar with our aims, and the remaining six were undergraduate students in linguistics who were also naive to the purposes of the experiment.

3.5.3 Dataset

Endoscopic recordings were kept under ten minutes to minimize participant discomfort and fatigue. In order to gather as much data as possible on the voicing contrasts themselves in the limited time available, distractor words were not used.

I elicited word-initial phonologically voiced and voiceless plosives and fricatives at the labial and alveolar places of articulation. Each obstruent was elicited in a real English word embedded in a carrier sentence, to control the phonetic context while providing the most natural mode of speech possible. The four obstruent voicing pairs /p b/, /t d/, /f v/, and /s z/ were elicited through minimal pairs. The words in their context sentences are shown in table 3.1.

This yielded data on two pairs of stops and two pairs of fricatives; two pairs of labials and two pairs of alveolars. All the contrasts are word-initial, as glottal abduction gestures are more pronounced initially than in other positions in normal speech (cf aspiration).

The frame sentence used was "Say *x* again." Table 3.1 gives the orthographic presentation form of the sentences used.

Say PEER again.
Say BEER again.
Say TIER again.
Say DEAR again.
Say FEAR again.
Say VEER again.
Say SEAL again.
Say ZEAL again.

Table 3.1: Orthographic presentation form of the sentences eliciting voicing contrasts.

All sentences have the same phonetic frame for the consonant being observed: [seɪCiɹəgɛn]. (The exception is that [l]-final words are used for /s z/ because no [ɹ]-final minimal pair is available in English.) For each sentence, primary phrasal stress fell on the target word; secondary stress on the final syllable of the sentence.

Two further sets of sentences illustrating a prosodic contrast were elicited after recording the above sentences, but will not be analysed in the current study. They include sentences with contrastive nuclear accent on the target word ("Say PEER again, not BEER again.") and sentences where contrastive nuclear accent falls immediately before the target word ("SAY peer again, don't WRITE peer again.") The full set of sentences recorded from each speaker is given in Appendix B.

3.5.4 Recordings

The equipment was sterilised before each recording. The speaker was seated in front of the microphone in a quiet room. Because the equipment was used internally, a medical doctor was present to supervise its use. This doctor was paid CAN$100 per hour for his time.

During placement, the endoscope was guided by the supervising medical doctor, with frequent checks that the participant was not experiencing undue discomfort.

The endoscope was guided through the nasal sinus along the medial edge of the inferior meatus. The two nasal cavities are not completely symmetrical in most people. Whether we used the left or right side of the nasal septum (left or right nostril) depended on the participant's preference and on the relative navigability. For two speakers, the first attempt was unsuccessful but we were able to use the other side.

When the lens of the endoscope was in place above the larynx, the participant was instructed to hold the tube in place where it entered the nose. Having control of the instrument inside them gave participants some reassurance. The duration of the insertion procedure varied greatly, from under two minutes to as long as three or four minutes, depending of the navigability of the nasal passages and the participants' level of comfort.

Eight index cards were used—one for each sentence in table 3.1. They were shuffled before each speaker's recording. The author presented the cards one at a time to the speaker. The speaker read each sentence twice in normal speech, then twice in whispered speech, with a pause after each repetition, before being shown the next card. The other two sets of utterances (illustrating an accented/post-accented contrast) were presented next in the same way.

The experimenters (the author and John Esling) attempted to keep the camera roughly centred on the glottal opening. Interruptions in the recording occurred for various reasons. The endoscope was occasionally displaced by movements of the velum and sometimes of the tongue and had to be moved back into position. A speaker's breath would sometimes fog up the lens. This could be cleared by having the speaker swallow—the associated tongue root retraction wiped the lens clear. Occasionally, the lens would drop or the epiglottis rise so that the two touched, causing a mild gag reaction. No subject was so discomforted at this stage that they asked to discontinue the recording, but a pause was often required to regain composure. Note that these events could cause the lens to end up higher or lower in the pharynx than before, meaning that lens-to-glottis distance varied even within a single recording (see section 2.5.2).

Recordings varied in duration from three minutes to ten minutes. In addition to the voicing contrasts and the prosodic dataset, further items were recorded if the speaker was comfortable—these included pitch manipulations on sung vowels, and other laryngeal modes (breathy voice, strong whisper). Only the voicing contrasts will be examined in this study.

3.5.5 Segmentation of the acoustic signal

Following data collection, the audio track was used to locate the consonants under investigation. In order to compare phonologically voiceless obstruents to phonologically voiced obstruents in normal speech, and to compare both to their counterparts in whispered speech, landmarks are needed that can be consistently identified in all conditions. The landmarks used for segmentation in this work (following the criteria outlined by Turk, Nakai & Sugahara 2006) are described below.

In wide-band spectrograms, the start and end of the fricatives were defined as the points where the formants of the surrounding vowels end and begin, as illustrated in figure 3.7.

The beginning of stop consonants was defined as the point where the formants of the preceding vowel ceased. The beginning of the release burst was used to mark the end of consonants. Aspiration is counted as part of the following vowel for two reasons: it is not present in voiced stops, and its whispered analogue, noted in the discussion of figure 3.5 above, is not reliably detectable by visual spectrographic means in whispered stops.

An example of the segmentation of [t] in normal and whispered speech is shown in figure 3.8.

3.5.6 Identifying frames to measure

Measurements of glottal aperture were made on individual frames. Video recordings have about 30 frames per second, or 1800 frames per minute, so in order to have a manageable amount of data, we must identify the frames of greatest interest for the current study.

The main experimental question is whether glottal aperture is different between phonologically voiced and phonologically voiceless segments in whispered speech,

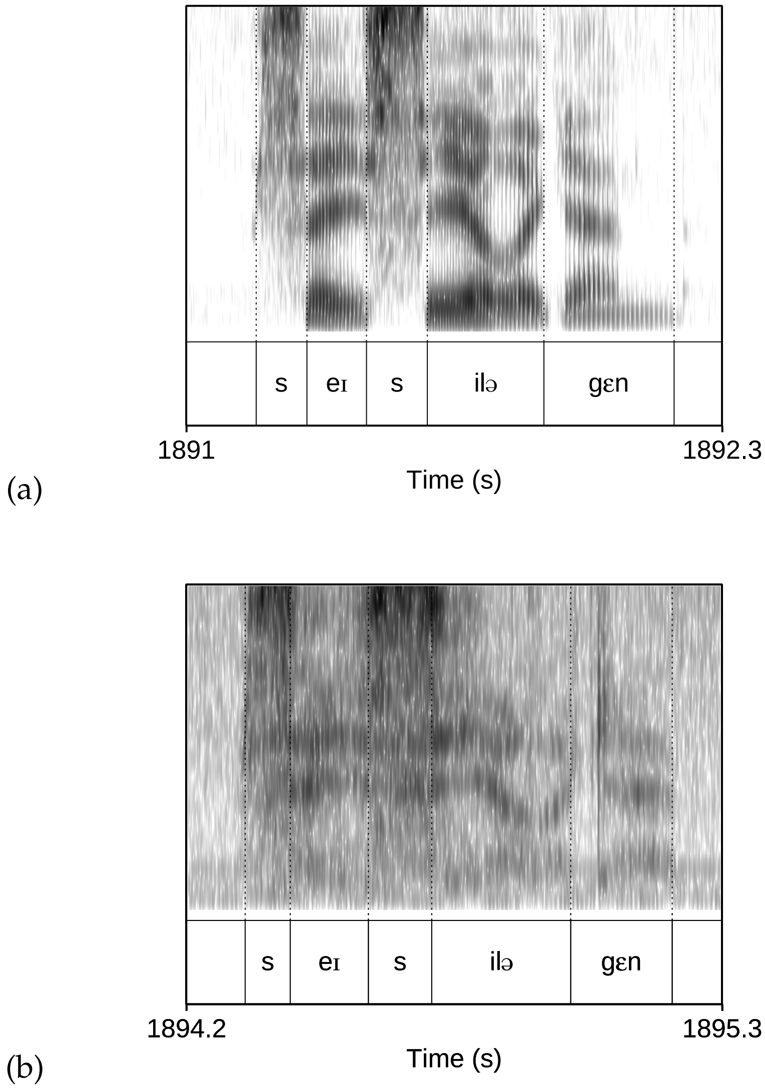


Figure 3.7: Spectrograms of "Say seal again" (a) spoken and (b) whispered. The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

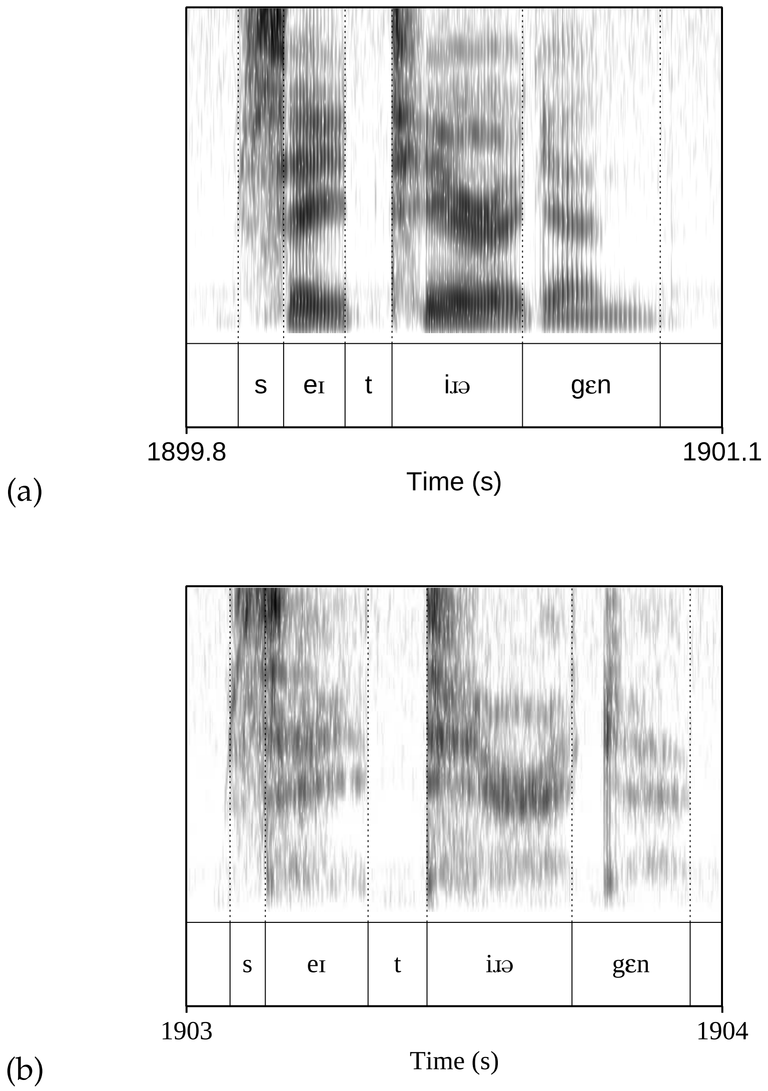


Figure 3.8: Spectrograms of “Say tier again” (a) spoken and (b) whispered. The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

as it is in normal speech. In order to examine this, we measured three data points per token: one before, one during, and one after the expected time of the abduction gesture. The frames for before and after the consonant are taken from the midpoints of the sonorant sequences preceding and following the target consonants ("v1" and "v2" in figures 3.9 through 3.12).

Two video processing utilities were tested: Adobe Premiere (Adobe 2001) and VirtualDub (Lee 2005)[2]. The main function required was the extraction of particular video frames, as determined by the temporal analysis of the audio track described above.

With each program, I observed the audio-visual synchronization at the onset of utterances following a non-speech interval at various points in each recording. In Adobe Premiere, I found video lags of up to six frames behind the audio track. This corresponds to between 150 and 190 ms, an unacceptable margin of error when looking at obstruent voicing. The glottal opening under investigation lasts for three to five frames, (see, for example, figures 3.9 and 3.11).

VirtualDub exhibited no detectable lag in any of the recordings, and so I used it to perform the required video-editing tasks.[3]

A subjective, visual inspection of the video of all speakers suggests that in normal speech, the maximum difference in glottal aperture between voiced and voiceless fricatives is reached during the frication. See for example the sequence of frames from a production of [s] in figure 3.9, compared with that for [z] in figure 3.10. While there is no abduction at all in the [z] token, there is a clear abduction gesture for [s]. It begins on frame 4, peaks on frame 6, and ends with full adduction restored on frame 9.

All tokens showed a glottal aperture peak between the onset and offset of frication. We could either select the frame with maximum glottal aperture, or select a frame based on an acoustic landmark. The problem with taking the frame with maximum glottal aperture is that, for voiced tokens such as that in figure 3.10,

[2] After measurements had been made, a third utility was found. Avidemux (Mean 2008), whose performance for the current tasks is comparable to that of VirtualDub, is also distributed under the GNU GPL license, and is available on a variety of operating systems. VirtualDub is only available for Windows.

[3] I used a very limited set of video-processing features: audio-video synchronization check, extraction of the audio signal, and extraction of still frames. This review bears only on those features. The two programs differ in many other ways, such as cost and nonlinear video editing functions. The foregoing implies no critique of their relative merits for other tasks.

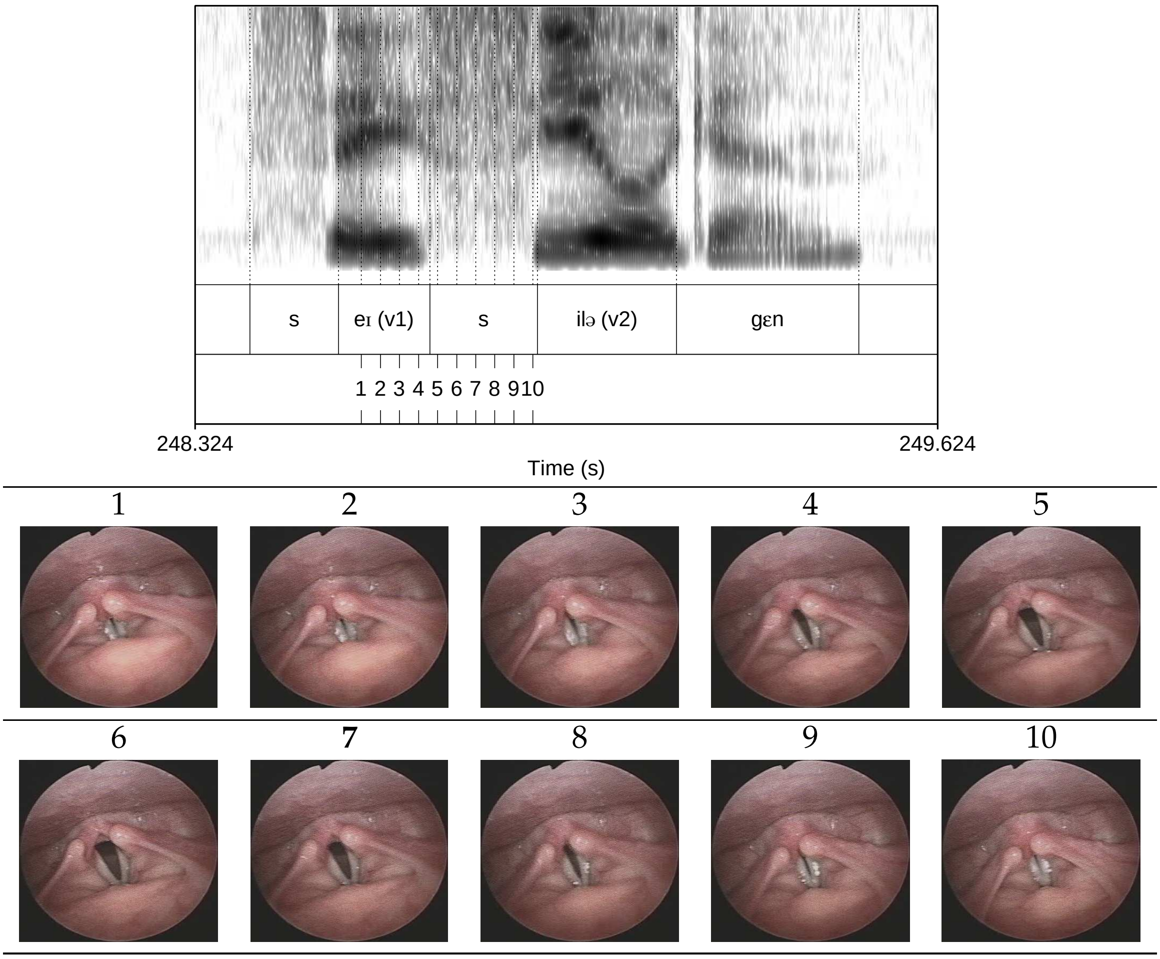


Figure 3.9: Frame sequence from spoken [s] illustrating glottal abduction. The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

there is no apparent difference in glottal aperture across the entire VCV sequence. In the interest of being able to compare conditions in as unbiased a fashion as possible, I simply measure the frame closest to the temporal midpoint of the fricative (as determined by the segmentation described in section 3.5.5 above. This frame (frame 7 in figure 3.9, frame 6 in figure 3.10) always showed substantial abduction for phonologically voiceless fricatives in normal speech. The maximum sometimes occurred earlier (as in figure 3.9) or later.

The maximum aperture for voiceless plosives is attained close to the release. Figure 3.11 illustrates this for [p], compared against [b] in figure 3.12. Again, the voiced segment shows no change in glottal aperture. The voiceless segment shows abduction starting on frame 4, peaking on frame 6, and ending by frame 8.

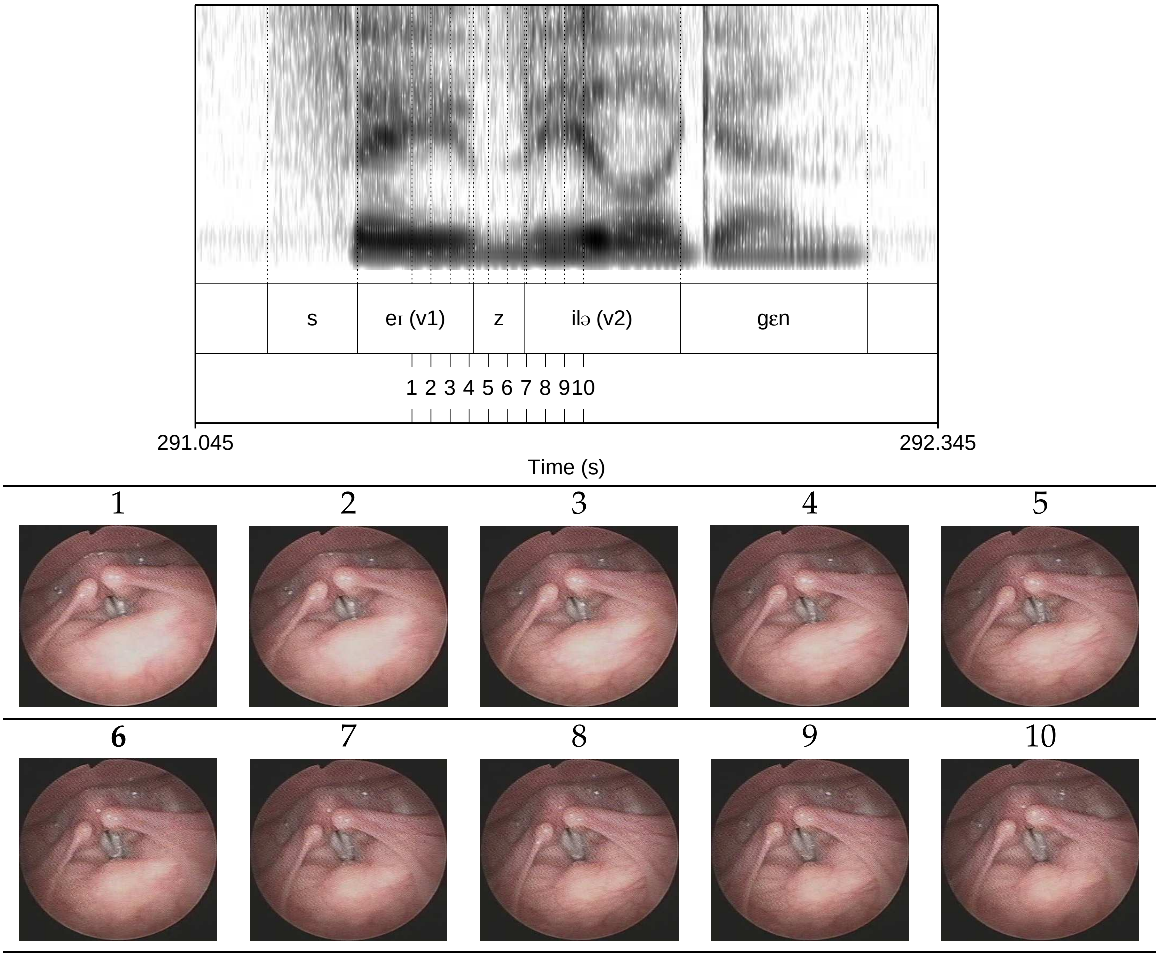


Figure 3.10: Frame sequence from spoken [z] illustrating lack of glottal abduction. The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

The frame closest to the point of release (frame 6 in figures 3.11 and 3.12) was selected for measurement in the plosives.

3.5.7 Extraction of corresponding video frame

Knowing the time at which we want to measure glottal aperture, the video frame closest to it can be isolated as a still image. Time points for desired frames were translated into (fractional) frame numbers based on 29.97 frames per second and a first frame synchronous with the first acoustic sample.[4] Equation 3.1 gives the

[4] More precisely, one field is produced every $\frac{1.001}{60}$ seconds (59.94 fields per second) (Kiver 1964, p320—NTSC technical specification F.2.). A field contains half the information for a full video frame—one field contains all the odd-numbered horizontal lines of pixels; the next field contains all the even-numbered lines. A full frame, therefore, is produced every $\frac{1.001}{30}$ seconds, or about 29.97 frames per second. Interested readers are referred to Kiver (1964) for more details on the historical and engineering justifications for these properties of the NTSC standard.

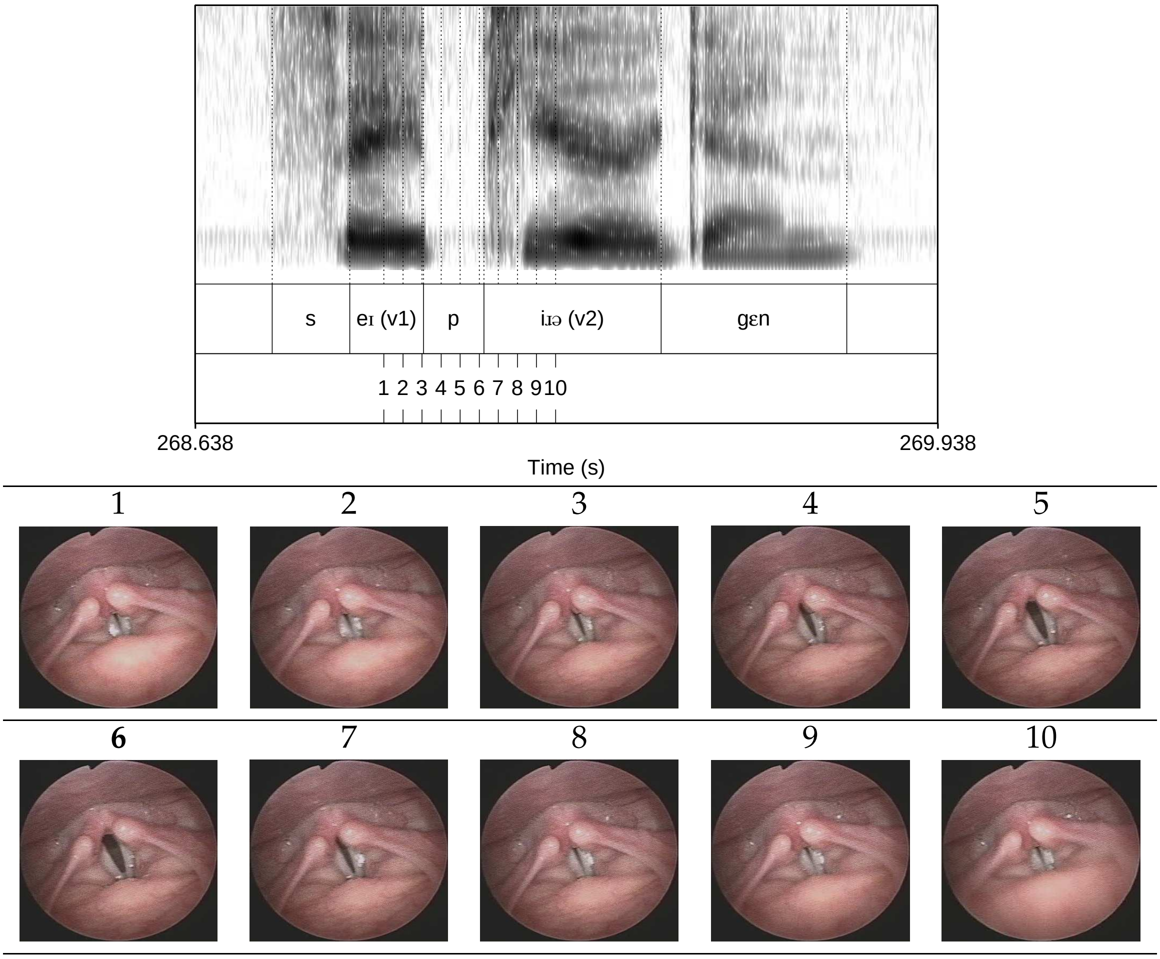


Figure 3.11: Frame sequences from spoken [p] illustrating glottal abduction. The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

calculation used to convert time (t) to a (fractional) frame number.

$$\text{frame} = t \cdot 29.97 \tag{3.1}$$

The calculated frame number was rounded to the nearest whole number (there are no part-numbered frames). This frame was isolated for analysis. Note that equation 3.1 assumes that video frame zero and audio sample zero are exactly synchronous. This property is satisfied in VirtualDub.

Because the visual data was recorded at 29.97 frames per second, the nearest frame to an identified acoustic landmark could be up to 16.7 ms away (mean ex-

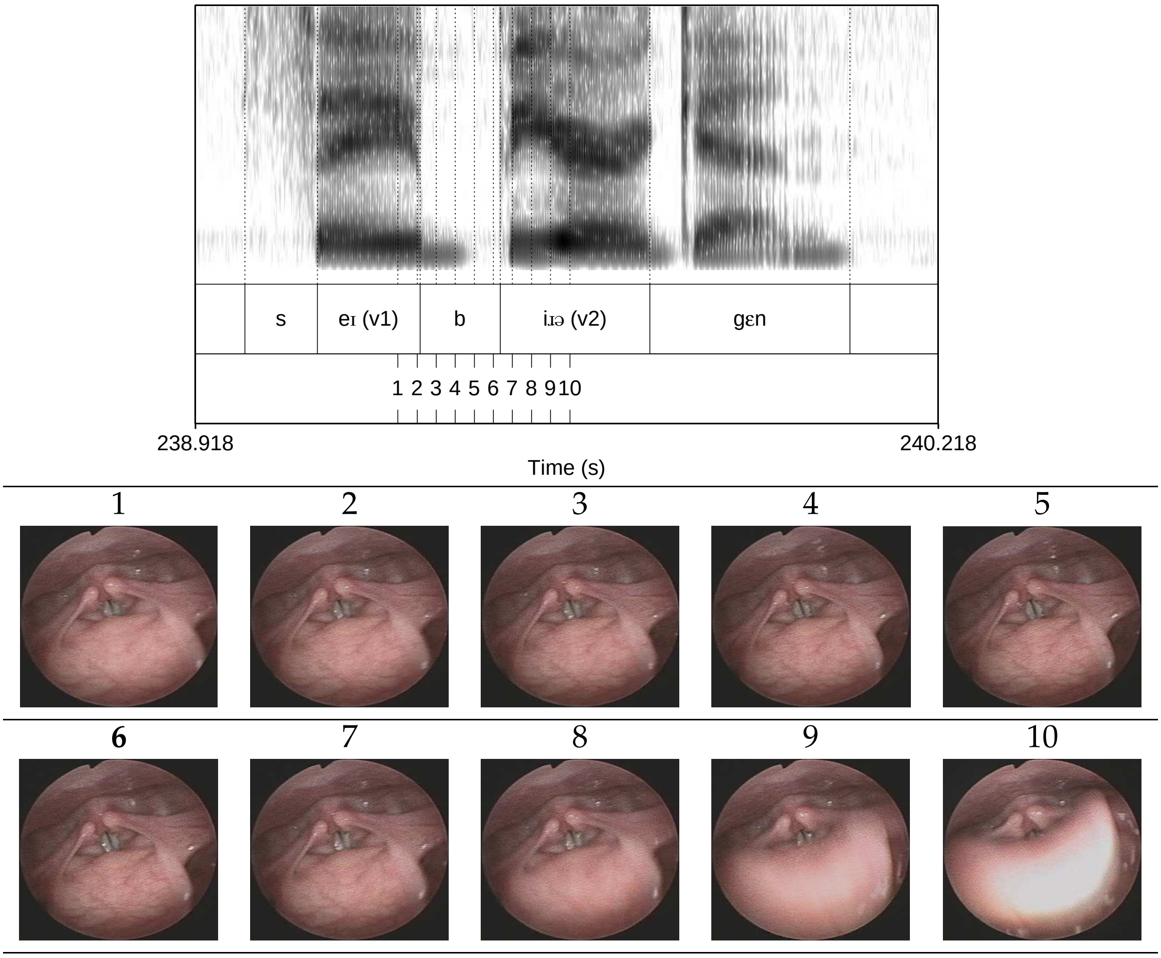


Figure 3.12: Frame sequence from spoken [b] illustrating lack of glottal abduction. The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

pected distance = 8.3 ms).[5] This is a small error compared to a mean duration of 135 ms for phonologically voiceless consonants across these recordings. Note also that the abduction gestures observed in figures 3.9 and 3.11 span five and four frames (respectively). With these facts in mind, a mismatch between identified acoustic landmark and measured video frame of no more than 16.7 ms is deemed acceptable.

[5] With a sampling rate of 30/1.001 frames per second (approximately 29.97), there are 33.3 ms between samples. Thus, the greatest distance one can be from the nearest sample (the maximum error) is 16.7 ms. Within the bounds, the magnitude of error should be a flat distribution; an error of 16.2 ms is as likely as an error of 0.3 ms. Thus, the mean error is half the maximum error, or about 8.3 ms. Alternatively, Peter Bell (pers.comm.) has calculated the standard deviation (keeping in mind that this is a flat, not a normal, distribution). The variance is $\frac{1}{P}\int_{-P/2}^{P/2} x^2 dx$, which works out to $\frac{P^2}{12}$; so the standard deviation is $\frac{P}{\sqrt{12}}$. For a period P between video frames of 33.3 ms, this works out to a standard deviation of 9.6 ms.

3.5.8 Measurement

With the frames to measure identified as described above, the procedure described in chapter 2 was used to gather sound quantitative measures of glottal aperture.

3.5.9 Annotation reliability test

A retest was performed to determine the extent to which annotator's judgments of the visual landmarks used in measurement were replicable.

Krippendorff (1980) identifies three types of reliability that can be tested: stability, reproducibility, and accuracy. "*Stability* is the degree to which a process is invariant or unchanging over time. . . . *Reproducibility* is the degree to which a process can be recreated under varying circumstances, at different locations, using different coders. . . . *Accuracy* is the degree to which a process functionally conforms to a known standard, or yields what it is designed to yield." (pp 130–131) In the current work, I was unable to use separate coders to check my measurements against. Nor do I have a known standard against which to test my results. Therefore, only stability is tested here—the extent to which my own judgments are consistent from one annotation session to another, separated by a suitable time break. Krippendorff calls this a *test-retest* design.

I performed a second set of measurements on a subset of the tokens and compared them to the first. The second set of measurements was performed 6 months after the first, without reference to the first measurements taken.

There were 292 tokens which were measurable—that is, on which glottal aperture and at least one cuneiform tubercle could be distinguished. Remember that for each token, three frames were identified for measurement: one before the consonant, one during the consonant, and one after the consonant. I randomly selected 30 video frames from each position (before, during, and after) for remeasurement, yielding a second measurement on slightly over ten percent of the measured frames.

In order to determine whether the second measurement of the validation set of 90 frames was close enough to the first, I ran statistical tests representative of those in the main analysis below (section 3.6). In this section I only report significance levels and directions of effects; for graphical presentation of the means and variances, see the full analysis of the data.

First, the vowels were analysed for the overall correlation between glottal aperture and mean cuneiform tubercle width, and for the difference between voiced and whispered phonation modes. The potential value of the cuneiform tubercle width measure as a proxy for camera-to-glottis distance was verified with a simple correlation: for both sets of measurements, glottal aperture and cuneiform tubercle width were significantly correlated. In the first set of measurements, the slope of the correlation was 0.15 (an increase of 1 pixel in cuneiform tubercle width corresponds to an increase of 0.15 pixels in glottal aperture), and $r^2 = 0.18$. In the second set of measurements, the slope of correlation was 0.17, and $r^2 = 0.20$. In both cases, the correlation was significant, with $p < 0.001$.

A linear mixed-effects model was fitted to the data[6] with glottal aperture as the dependent (response) variable, mean cuneiform tubercle width as a covariate, phonation mode (voiced/whispered) as a fixed factor, and speaker as a random factor. The first set of measurements showed no significant correlation between glottal aperture and cuneiform tubercle width in this model (slope=0.012; p=0.768). There was a significant effect of phonation mode, with glottal apertures estimated at 15 pixels greater in whispered than in voiced vowels (p=0.011). The second set of measurements gave similar results for the correlation (slope=0.057; p=0.5084) and for phonation mode (diff=16 px; p=0.016). The same inferences would be drawn for either set of measurements: there is no significant linear relationship between measured glottal aperture and measured cuneiform tubercle width in vowels; and whispered vowels have a greater glottal aperture than voiced vowels.

Second, the consonants were examined. A linear mixed-effects model was fitted to the data with mean cuneiform tubercle width as a covariate, phonation mode (spoken or whispered) and phonological voicing (voiced or voiceless) as fixed factors, and speaker as a random factor. Neither the first nor the second set of

[6] See section 3.5.11 below for more on why linear mixed-effects models, rather than more traditional ANOVAs, were used with the current data. All significance values are based on the Monte-Carlo Markov Chain sampling method (10 000 samples) used in the `lme4` package in R (R 2008).

measurements showed any significant effects or interactions with this model. Because of the ambiguous correlation in the vowel analysis, a second model was fitted without the cuneiform tubercle width as a covariate (but otherwise identical). For the first set of measurements, this new model yielded no effect of phonation mode (p=0.305), but a significant effect of voicing (p=0.008): glottal apertures of phonologically voiceless consonants averaged 22 pixels greater than those of phonologically voiced consonants. There was no significant interaction between phonation mode and phonological voicing (p=0.631). The second set of measurements yielded the same pattern: no effect of phonation mode (p=0.347), a significant effect of phonological voicing (p=0.002, magnitude=26 pixels), and no interaction (p=0.761). The same inference would be drawn for either set of measurements. There is no significant linear relationship between measured glottal aperture and measured cuneiform tubercle width in consonants. The overall phonation mode of an utterance (whispered or normal speech) has no effect on the glottal aperture of consonants, but the phonological identity—voiced or voiceless—has a strong effect in both normal and whispered speech.

The above tests establish that the measurement procedure followed in this study is stable between measurements by the same annotator. We can therefore have reasonable confidence in the validity of the results reported below.

3.5.10 Measurement—summary

Following is a step-by-step summary of the procedures followed in measuring data for this study. The first sequence describes the procedure for obtaining undistorted images from the video data.

1. Segment acoustic signal to determine time points for which to measure frames. (see section 3.5.5)
2. Convert time points to frame numbers: multiply by frame rate, in this case NTSC 29.97 frames per second, and round to nearest whole frame. (see section 3.5.6)
3. Extract frames: With the video open in VirtualDub, select File > Save image sequence..., and save to JPEG format (full quality); with the video open in Avidemux, select "File > Save > Save Selection as JPEG Images". This generates a directory with a file for each frame in the video, numbered. Select the ones you wish to measure. (see section 3.5.7)

4. Correct rectangular pixels. (Test magnitude of distortion first by comparing vertical size of "circular" field of view to horizontal size.) In this case, increase vertical dimension by 10%. (see section 2.4.1)
5. Measure and correct resulting image for barrel distortion. (see sections 2.4.2 and 2.4.3)

The second sequence describes the measurements made on the images after distortion was removed (based on section 2.5).

1. Mark the size of each cuneiform tubercle. Use the maximum visible width to ensure consistency across tokens within a speaker.
2. Mark the vocal folds with straight-line approximations between the visible endpoints. Record the length in pixels and the angle of each vocal fold.
3. Calculate the angle of the glottal aperture. It should be perpendicular to the angle bisecting the two vocal folds (as depicted in the wireframe at the right of figure 2.10).
4. Measure the glottal aperture. It should be the maximum visible distance between the vocal folds, using a line at the angle calculated in the previous step.
5. For each token, calculate the obstruction-adjusted glottal aperture as the product of the measured glottal aperture and the length of the longer vocal fold length, divided by the shorter vocal fold length.

3.5.11 Statistics used

In this study and the one described in chapters 4 and 5, I use linear mixed effects models to analyse the data, and Markov chain Monte Carlo (MCMC) sampling to establish the significance levels of the results. Linear mixed effects models are designed to deal with data where some factors are fixed and others are random. MCMC methods allow inferences to be drawn that do not depend on normality or homogeneity of variance—conditions which are required for other methods used with this sort of data, such as Repeated Measures ANOVA.

Fixed factors systematically sample many or all relevant levels; they can be measured at the same levels in separate studies. Phonation mode is a fixed factor—a replication of the current study could include both levels of this variable (normal and whisper) just as they are used here. *Random factors* cannot be measured at the same levels in separate studies, because any one study includes only a random sample of levels of the factor. The primary example of a random factor is participant. A replication of the current study would not use the same participants. Random factors tend to represent large populations, such as speakers or words.

MCMC estimation of significance levels uses the data itself to determine a "posterior distribution" against which the null and alternative hypotheses are evaluated. By basing its estimate on the distribution of the data being examined, without reference to externally-defined reference distributions (such as the normal distribution), the MCMC technique liberates the analyst from having to check normality, or even homogeneity of variances. MCMC is also less affected by small sample sets than other techniques are.

Except where noted otherwise, statistical tests in this chapter employ a linear mixed-effects model with phonation mode (normal or whispered) and phonological voicing (voiced or voiceless) as fixed factors and speaker (9 levels) as a random factor. The dependent variable (generally either plain or adjusted glottal aperture on a linear or logarithmic scale) is noted in each reported result.

Following Baayen (2008, pp 241–259) and Quené & van den Bergh (2008), I use the *lmer* (Linear Mixed Effect Regression) function from the `lme4` package (Bates 2005) in R (R 2008). I report p-values derived from Markov Chain Monte Carlo (MCMC) sampling (10 000 samples). For further discussion of the theory behind mixed-effects models, see Neter, Kutner, Nachtsheim & Wasserman (1996), Bates (2005), and Baayen (2008, chapter 7).

Because I am using MCMC sampling to determine p-values, there are no F-statistics to report. I include graphs showing group variances from the data to give the reader a visual idea of the effects reported. In order to avoid inflation of the variances due to inter-speaker differences, a manipulation was performed akin to the statistical manipulation used to exclude inter-level variation from the random variables.

Where indicated, I include cuneiform tubercle width as a covariate in the model.

3.6 Results

I begin by presenting some patterns observed in the glottal aperture of vowels, which support the inclusion of cuneiform tubercle width as a covariate measure.

I also compare the plain glottal aperture measure to the obstruction-adjusted glottal aperture measure. I then examine the data for the consonants themselves.

3.6.1 Controls and confounds

The principal analysis of this study is performed on the consonantal measures. I therefore chose to use the measured vowels as a semi-independent dataset to validate the control and confound measurements. I begin, in section 3.6.2, by exploring the option of log-scaling the glottal aperture and cuneiform tubercle measures to approximate a normal distribution. In section 3.6.3, I evaluate the use of cuneiform tubercle width as a covariate standing in for camera-to-glottis distance. In section 3.6.4, I examine the benefit of using the ratio of visible vocal fold lengths to adjust for obstruction by cuneiform tubercles.

3.6.2 Distributions and log scaling

Figure 3.13 shows the distribution of glottal aperture values across all vowels produced. While the difference between spoken (black ’o’) and whispered (red ’+’) glottal apertures is clear, it is also clear that variance increases as the measure increases.

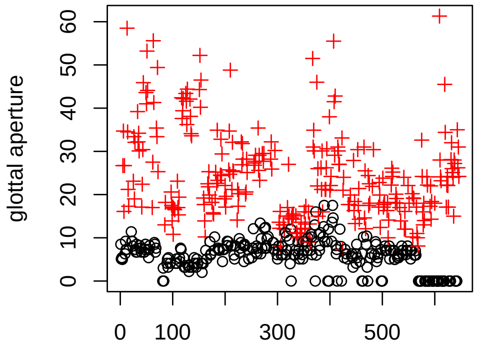


Figure 3.13: Glottal aperture for all spoken (black ’o’) and whispered (red ’+’) vowels produced in the current data. The horizontal axis is the order of production.

The same tendency is evident when we look at the variance of the two groups—spoken and whispered—in the error bars shown in figure 3.14. The group with the higher mean (whispered vowels) also has greater variance (about 2.9 times as much as spoken vowels).

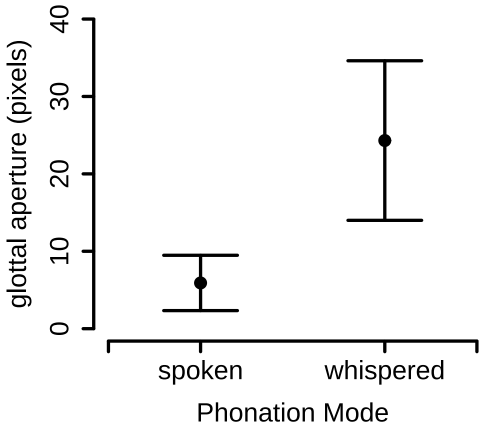


Figure 3.14: Glottal aperture error bars for spoken and whispered vowels. The centre dot is the mean for each group; the bars represent one standard deviation above and below the mean.

A similarly asymmetrical variance, though slightly less pronounced, is evident in the cuneiform tubercle measures, shown in figure 3.15. Figure 3.16 shows the

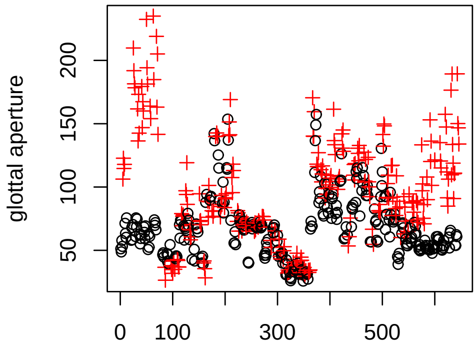


Figure 3.15: Cuneiform tubercle width for all spoken (black 'o') and whispered (red '+') vowels produced in the current data. The horizontal axis is the order of production.

error bars for this data. The cuneiform tubercle width of whispered vowels has a standard deviation 1.8 times that of spoken vowels.

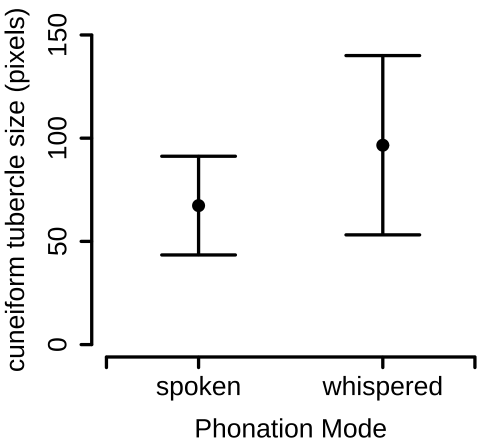


Figure 3.16: Cuneiform tubercle size error bars for spoken and whispered vowels. The centre dot is the mean for each group; the bars represent one standard deviation above and below the mean.

This pattern suggests log-transforming the data, so that variances remain similar across categories with different means (such as the spoken and whispered vowels in figure 3.13). However, there are several zero values for glottal aperture. Log-transforming zero yields $-\infty$, which renders meaningful statistical analyses impossible. Therefore, before transforming, all measurements were increased by 0.001 pixels. Such a small shift does not greatly effect the transformed value for non-zero measurements, but it makes the zero measurements finite once transformed. (Note that the logarithmic transform will always preserve relative magnitudes—if $a > b$ then $\log a > \log b$—and this important property is unaffected by the adjustment.) Figure 3.17 shows the glottal aperture measures log-transformed (base e); figure 3.18 shows the cuneiform tubercle measures log-transformed.

For both the glottal aperture and the cuneiform tubercle width measures, the transform has solved the problem of greater spread at greater values. However, note that the originally-zero measurements, even after transformation, stand well apart from the rest of the measurements. This is a problem, as clearly shown by the error bars in figure 3.19. The presence of the zeros causes the variance of the spoken vowels to be very wide.

If the zeros are omitted, the variance of the remaining spoken vowels is seen to be similar to that for whispered vowels (figure 3.20).

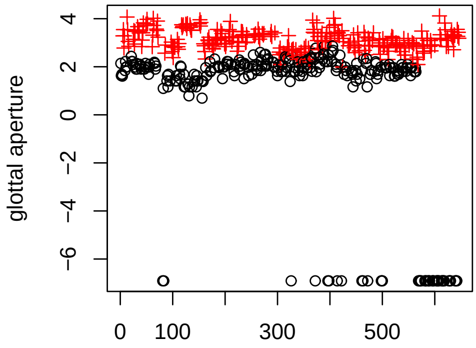


Figure 3.17: Log-scaled glottal aperture for all spoken (black 'o') and whispered (red '+') vowels produced in the current data. The horizontal axis is the order of production.

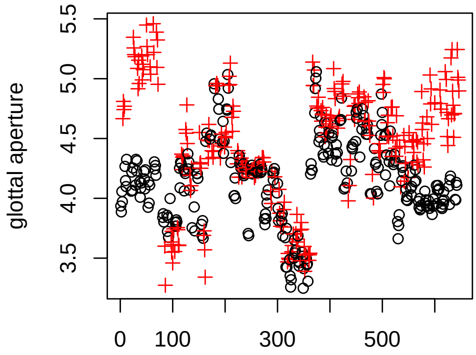


Figure 3.18: Log-scaled cuneiform tubercle width for all spoken (black 'o') and whispered (red '+') vowels produced in the current data. The horizontal axis is the order of production.

3.6.3 Cuneiform tubercle width as a covariate

In figure 3.21, glottal aperture and cuneiform tubercle width are compared for all recorded vowels (n=565). Using a mixed-effects linear model with glottal aperture as a dependent variable and cuneiform tubercle width as a covariate, phonation mode (normal vs whispered) as a fixed factor, and speaker as a random factor, we see that there is indeed a significant effect of phonation mode

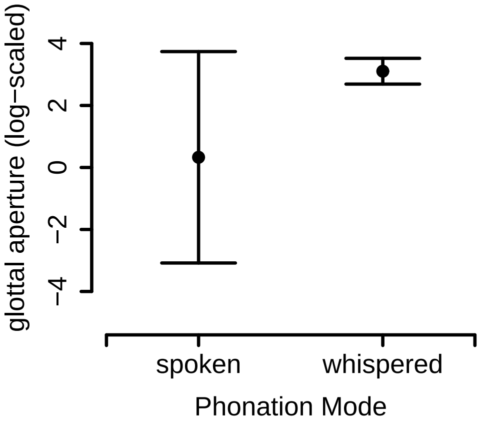


Figure 3.19: Log-scaled glottal aperture error bars for spoken and whispered vowels. The centre dot is the mean for each group; the bars represent one standard deviation above and below the mean.

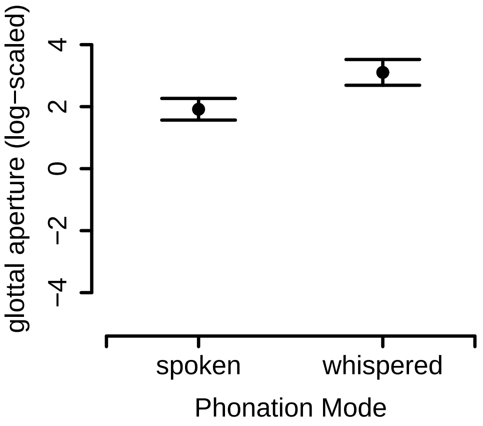


Figure 3.20: Log-scaled glottal aperture error bars for spoken and whispered vowels (zeros omitted). The centre dot is the mean for each group; the bars represent one standard deviation above and below the mean.

($p < 0.001$)—glottal aperture for whispered vowels averages 16 pixels greater than for voiced vowels. Cuneiform tubercle width does not show a significant correlation with glottal aperture ($p = 0.125$); however, there is a significant interaction with phonation mode ($p < 0.001$). That is, there is no *overall* trend, but the difference in regression slopes between spoken (slope$= 0.026$) and whispered (slope$= 0.126$) vowels *is* significant. In separate post hoc tests on the whispered and spoken subsets of the vowel data, the correlation between glottal aperture and cuneiform tubercle width is significant for whispered vowels ($p < 0.001$, n=276), but not for spoken vowels ($p = 0.704$, n=289).

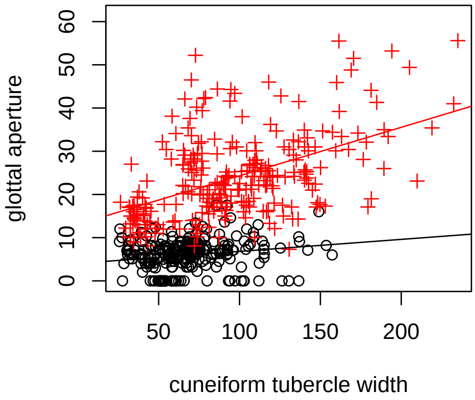


Figure 3.21: Glottal aperture plotted against mean cuneiform tubercle width for spoken (black 'o') and whispered (red '+') vowels. Regression lines are shown for each phonation type (spoken = black, whispered = red).

The smaller slope of correlation in spoken vowels (0.026) than in whispered vowels (0.126) is due to the fact that the spoken vowels are at the floor of the range—they all represent adducted vocal folds vibrating—and so glottal aperture is not as free to vary as it is in the more abducted whispered vowels. In particular, note the several spoken vowels with zero glottal aperture at the bottom of the graph in figure 3.21, lying along most of the range of the cuneiform tubercle width measure.

The significant difference in glottal aperture between spoken and whispered vowels remains if we use a linear model without cuneiform tubercle width as a covariate ($p < 0.001$). However, this model only accounts for 67% of the variation in glottal aperture ($R^2 = 0.666$). The linear model incorporating cuneiform tubercle width explains 76% of the variation ($R^2 = 0.756$)—a substantial gain over the simpler model.

I ran a separate mixed-effect model with cuneiform tubercle width as dependent variable, phonation mode as a fixed factor, and speaker as a random factor to determine whether there is a systematic difference in cuneiform tubercle width between spoken and whispered vowels. There is a significant effect of phonation mode ($p < 0.001$): spoken vowels have a mean cuneiform tubercle width of 68.07 pixels; whispered vowels average 96.77 pixels. The still images in figure 3.22 illustrate the general pattern that whispered vowels differ from spoken vowels in both laryngeal height and glottal opening. We can infer that whispered vowels tend to be closer to the camera than spoken vowels (larynx raising). This is consistent with the observations of Esling (2002) and Esling & Harris (2003) on laryngeal sphinctering, in which the contraction of laryngeal structures (sphinctering) in whisper is accompanied by apparent laryngeal raising.

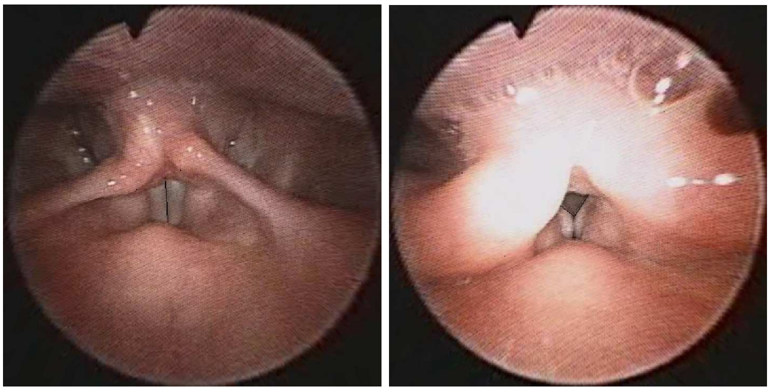


Figure 3.22: Video frames of spoken and whispered vowels, from speaker m1.

The above establishes the cuneiform tubercle width as a potentially useful covariate to include in statistical models of glottal aperture. It also shows that this usefulness can be diminished when glottal aperture variation is constrained by a floor effect. As a result, I use cuneiform tubercle width as a covariate in analyses below, but if an initial model shows that it contributes little, I exclude it from secondary analyses.

3.6.4 *Adjustment for cuneiform tubercle obstruction*

Figure 3.23 re-presents figure 3.21 using the obstruction-adjusted glottal aperture measure (designed to remove the confounding influence of one cuneiform tubercle partially obstructing a vocal fold, and thus reducing the apparent glottal aperture). Notice that the variance has increased relative to the trends observed (as expected). The overall increase in range of the adjusted glottal aperture measure in figure 3.23 over the plain glottal aperture measure in figure 3.21 is because equation 2.5 multiplies the plain glottal aperture by a factor of no less than 1 to obtain the adjusted glottal aperture measure. This increase in itself is not important, as the units (pixels) are already arbitrary relative to actual distances. There is still a strong main effect of phonation mode ($p < 0.001$). Using the obstruction-adjusted glottal aperture measure, we find a significant correlation

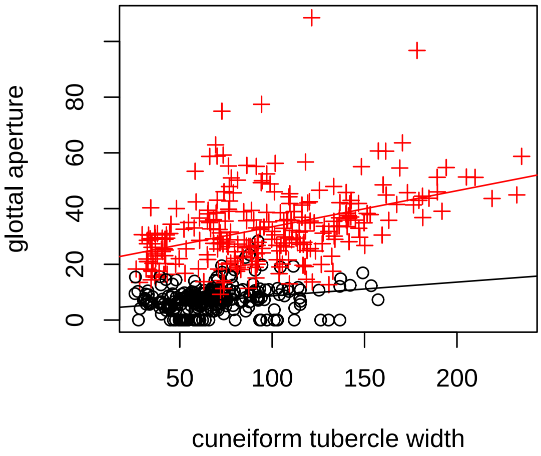


Figure 3.23: Adjusted glottal aperture plotted against mean cuneiform tubercle width for spoken (black 'o') and whispered (red '+') vowels. Regression lines are shown for each phonation type (spoken = black, whispered = red).

with cuneiform tubercle width ($p = 0.0083$), which interacts significantly with phonation mode ($p = 0.002$)—the slope is 0.066 for spoken and 0.149 for whispered tokens. In separate post hoc tests on the whispered and spoken subsets, we find (as we did for the plain glottal aperture measure) that there is a significant correlation for the whispered vowels ($p < 0.001$) but not for the spoken vowels ($p = 0.596$).

The linear model with the plain glottal aperture measure accounts for 76% of the variation ($R^2 = 0.756$), the linear model with the adjusted glottal aperture measure accounts for 73% of the variation ($R^2 = 0.733$). The two are comparable, therefore, in their explanatory power. The main difference is that the adjusted glottal aperture measure exhibits a significant overall correlation with cuneiform tubercle width, while the plain glottal aperture measure does not. This appears to be due to the asymmetric effect of the glottal aperture adjustment on spoken and whispered tokens. See section 2.5.3 for details of this asymmetric effect.

3.6.5 Summary

The correlations between glottal aperture and cuneiform tubercle width (in conditions without floor effects) justify our inclusion of the latter as a covariate to control for the variations in lens-to-glottis distance. This is particularly important because the larynx is systematically higher in whispered tokens than in normal tokens; a failure to control for larynx height could bias our results. The greater larynx height in whisper, indicated by the systematically larger cuneiform tubercle width, is probably due to the overall laryngeal sphinctering that accompanies whisper, reported by Esling & Harris (2005). Without accounting for this, we would infer that glottal apertures are different for phonologically voiced consonants in whisper than they are in normal speech. A further argument for inclusion of the covariate is that it yields a linear model which accounts for an extra 9% of the overall variance in glottal aperture.

While the adjustment of glottal aperture to correct for partial cuneiform tubercle obstruction of the glottis (described in section 2.5.3) does not seem to increase the precision of our results (there is no improvement in the proportion of variation accounted for by measured factors), it does seem to affect the results. The correction modifies whispered tokens more strongly than it modifies spoken tokens, and subjective visual observation of the measured frames suggests that cuneiform tubercle obstruction of a vocal fold is substantially more common in whispered speech than in normal speech.

It is reassuring that our results are the same for the obstruction-adjusted as for the plain glottal aperture measure. The only point of disagreement is that we found a significant correlation between the *adjusted* glottal aperture and cuneiform tubercle width measures in vowels, but no significant correlation between *plain* glottal aperture and cuneiform tubercle width in vowels. Because it agrees with the geometric prediction that apparent glottal aperture should be correlated with apparent cuneiform tubercle size, this difference tends to support of the adjusted measure over the plain measure.

However, this difference does not provide conclusive validation for the correction. We have no independent confirmation that apparent cuneiform tubercle width is directly correlated with larynx height, for example. Until better evidence either way is available, it is prudent to use both the plain and adjusted measures and rely on consensus between the results to justify strong conclusions in our analyses.

3.6.6 Consonants

Having established the validity of certain of our corrective measures using the vowel data, we can now turn to the consonants which are the object of this study.

I begin by presenting results for normal speech (section 3.6.7) and whispered speech (section 3.6.8) separately, before comparing the two (section 3.6.9).

3.6.7 Normal speech

Figure 3.24 shows a typical vowel-consonant-vowel sequence of three measured frames for the sentence "Say peer again" spoken aloud by a female participant. The glottal abduction gesture is plainly apparent in the middle frame (the consonant).

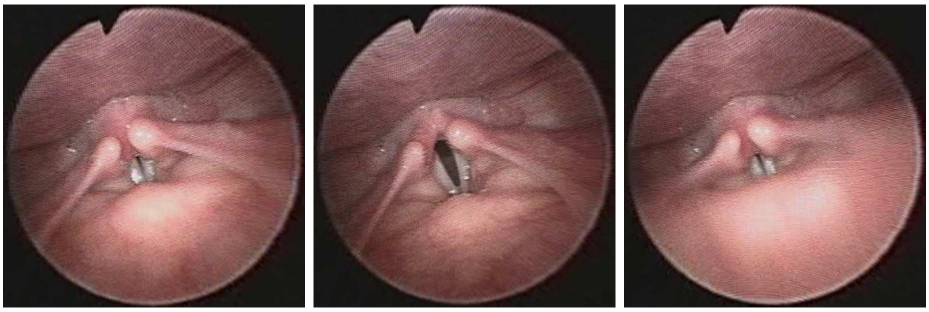


Figure 3.24: Typical VCV sequence of frames—"Say peer again" in normal speech by speaker f1.

Figure 3.25 shows a typical sequence for the sentence "Say beer again", spoken aloud by the same speaker. The absence of an abduction gesture is similarly clear.

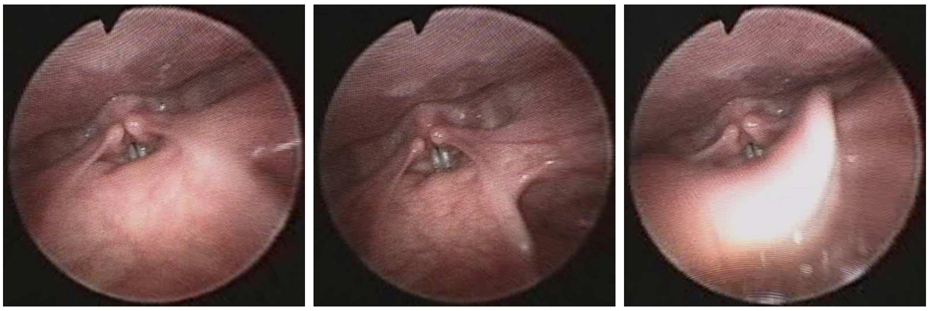


Figure 3.25: Typical VCV sequence of frames—"Say beer again" in normal speech by speaker f1.

This contrast is reflected in figure 3.26, which plots mean glottal aperture across the three measured points (preceding vowel, consonant, following vowel) for all spoken tokens, with error bars for each point. Note the greater variance in the voiceless consonants' glottal aperture. This greater variance at higher values of glottal aperture motivates log-scaling of the data (section 3.6.2 above), as shown

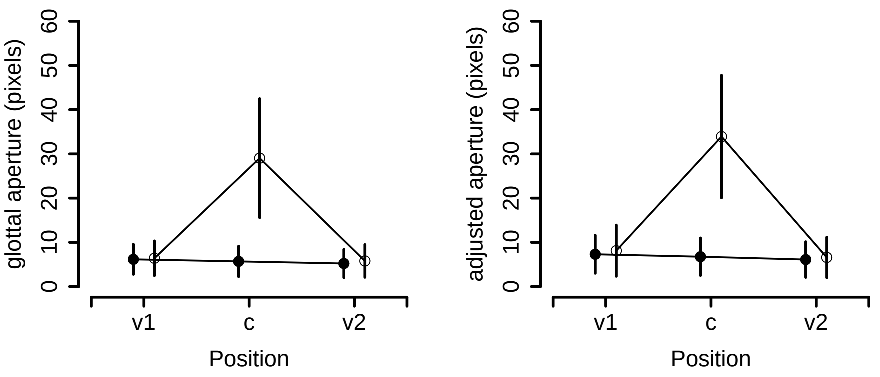


Figure 3.26: Glottal aperture plots (left=plain, right=adjusted for cuneiform tubercle obstruction) for spoken obstruents (averaged over all speakers). The points are taken from the vowel preceding the consonant (v1), the consonant itself (c), and the vocalic sequence following the consonant (v2). The filled circles represent voiced tokens, and the open circles represent voiceless tokens. Vertical lines represent one standard deviation above and below each mean.

in figure 3.27. Note, however, the high variance for the vowels and voiced consonants. This is due to the influence of the zero-valued tokens (as discussed in section 3.6.2). Omitting the zero-valued tokens, we get figure 3.28, in which variances are similar to each other.

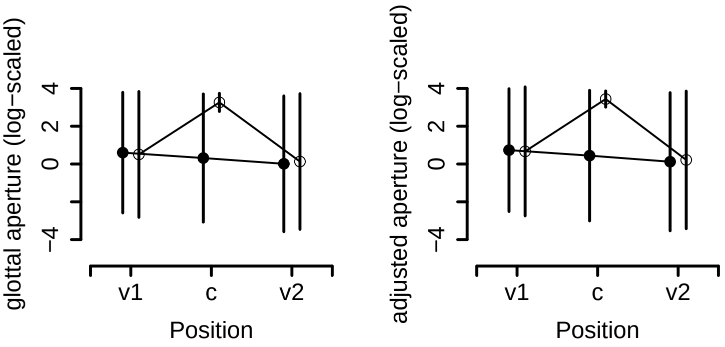


Figure 3.27: Log-scaled glottal aperture plots (left=plain, right=adjusted for cuneiform tubercle obstruction) for spoken obstruents (averaged over all speakers). The points are taken from the vowel preceding the consonant (v1), the consonant itself (c), and the vocalic sequence following the consonant (v2). The filled circles represent voiced tokens, and the open circles represent voiceless tokens. Vertical lines represent one standard deviation above and below each mean.

This mountain-and-plain pattern in spoken utterances is typical for all minimal pairs produced by all speakers, using both the plain and adjusted glottal aperture

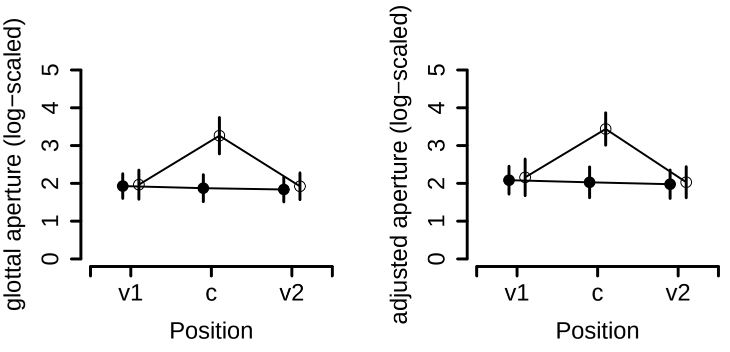


Figure 3.28: Log-scaled glottal aperture plots (left=plain, right=adjusted for cuneiform tubercle obstruction) for spoken obstruents (averaged over all speakers, zeros omitted). The points are taken from the vowel preceding the consonant (v1), the consonant itself (c), and the vocalic sequence following the consonant (v2). The filled circles represent voiced tokens, and the open circles represent voiceless tokens. Vertical lines represent one standard deviation above and below each mean.

measures, with three exceptions: two produced by speaker f5 and one produced by speaker m2.

The female speaker f5 produced the "tier"/"deer" contrast as shown in figure 3.29 and the "peer"/"beer" contrast as shown in figure 3.30, both of which diverge significantly from the overall pattern shown in figure 3.26.

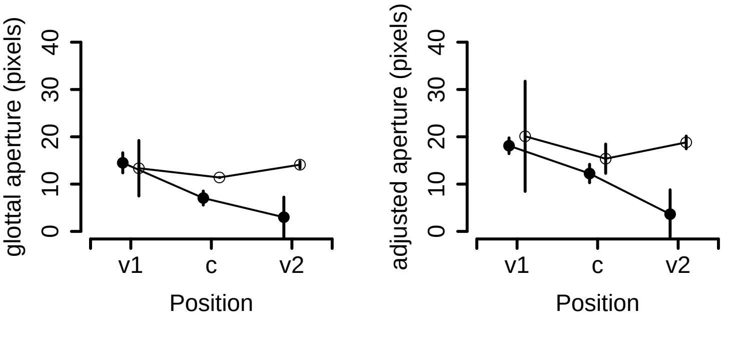


Figure 3.29: Glottal aperture plots (left=plain, right=adjusted for cuneiform tubercle obstruction) for spoken sentences "Say tier again" and "Say deer again" from female speaker f5. The points are taken from the vowel preceding the consonant (v1), the consonant itself (c), and the vocalic sequence following the consonant (v2). The filled circles represent voiced tokens, and the open circles represent voiceless tokens. Vertical lines represent one standard deviation above and below each mean.

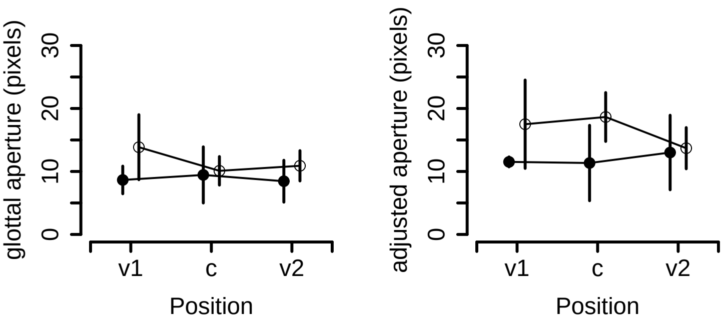


Figure 3.30: Glottal aperture plots (left=plain, right=adjusted for cuneiform tubercle obstruction) for spoken sentences "Say peer again" and "Say beer again" from female speaker f5. The points are taken from the vowel preceding the consonant (v1), the consonant itself (c), and the vocalic sequence following the consonant (v2). The filled circles represent voiced tokens, and the open circles represent voiceless tokens. Vertical lines represent one standard deviation above and below each mean.

The male speaker m2 produced the "fear"/"veer" contrast as shown in figure 3.31. This pattern is reminiscent of the mountain-and-plain pattern of figure 3.26, except that the "plain" is much higher.

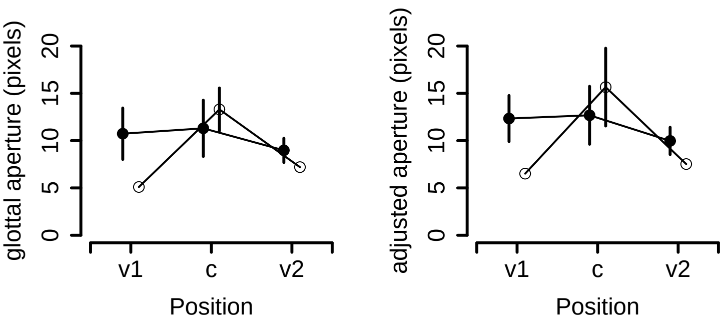


Figure 3.31: Glottal aperture plots (left=plain, right=adjusted for cuneiform tubercle obstruction) for spoken sentences "Say fear again" and "Say veer again" from male speaker m2. The points are taken from the vowel preceding the consonant (v1), the consonant itself (c), and the vocalic sequence following the consonant (v2). The filled circles represent voiced tokens, and the open circles represent voiceless tokens. Vertical lines represent one standard deviation above and below each mean.

Because each speaker only produced two spoken tokens of each target word, we cannot determine whether these are systematic or accidental exceptions, or whether they are part of the normal range of variation in glottal behaviour. These apparently-anomalous tokens are therefore not excluded from the dataset in the statistical analyses.

A mixed-effects linear model analysis run on the consonant data from normal speech, with glottal aperture as the dependent variable, cuneiform tubercle width as a covariate, phonological voicing as a fixed factor, and speaker as a random factor affirms that the "mountain-and-plain" pattern dominates, despite the anomalous pairs mentioned. There is a significant correlation between glottal aperture and cuneiform tubercle width ($p = 0.031$, adjusted glottal aperture $p = 0.042$). There is a significant effect of voicing: glottal aperture for phonologically voiceless consonants, at 29 pixels (px) is 23 px (411%) greater than for phonologically voiced consonants, at 6 px ($p < 0.001$ for both plain and adjusted glottal aperture measures). There is no interaction between the correlation of glottal aperture with cuneiform tubercle width and the effect of phonological voicing on glottal aperture ($p = 0.330$, adjusted $p = 0.093$).

3.6.8 Whispered speech

Figures 3.32 and 3.33 show sequences for whispered "peer" and "beer", respectively. The whispered [p] does not show the clear abduction we see in its spoken counterpart in figure 3.24. It seems to have the same glottal posture as the surrounding vowels. The whispered [b], on the other hand, does not (as its spoken counterpart, figure 3.25) show a constant glottal posture; we seem to see an *adduction* during the consonant!

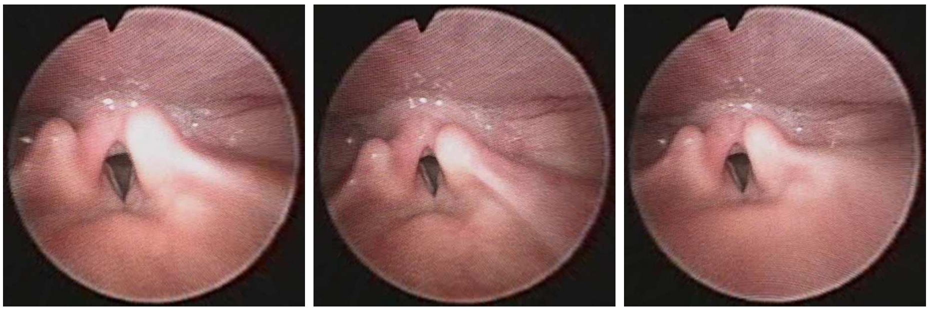


Figure 3.32: Typical VCV sequence of frames—"Say peer again" whispered by speaker f1.

This contrast is reflected in figure 3.34, which plots mean glottal aperture across the three measured points (preceding vowel, consonant, following vowel) for all spoken tokens, with error bars for each point.

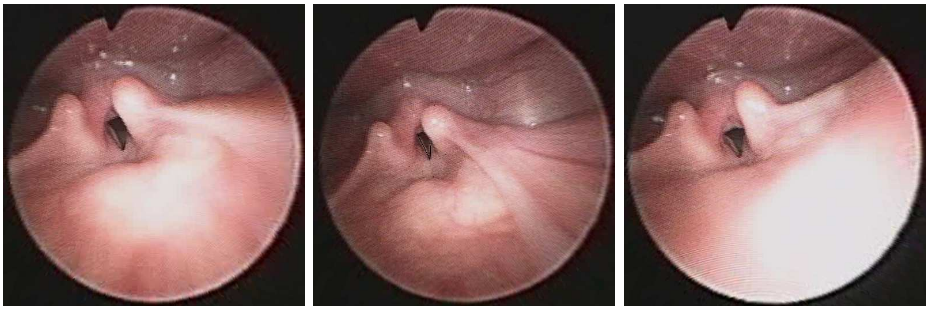


Figure 3.33: Typical VCV sequence of frames—"Say beer again" whispered by speaker f1.

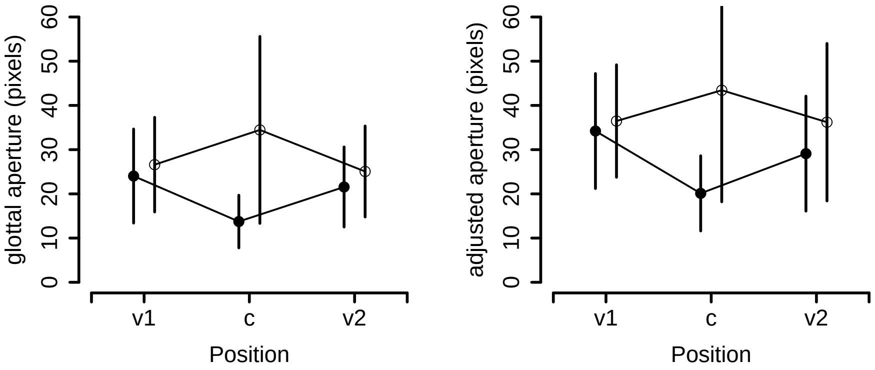


Figure 3.34: Glottal aperture plots (left=plain, right=adjusted for cuneiform tubercle obstruction) for whispered obstruents (averaged over all speakers). The points are taken from the vowel preceding the consonant (v1), the consonant itself (c), and the vocalic sequence following the consonant (v2). The filled circles represent voiced tokens, and the open circles represent voiceless tokens. Vertical lines represent one standard deviation above and below each mean.

3.6.9 Comparing normal to whispered speech

I ran a mixed-effects linear model analysis (separately for the plain and the adjusted glottal aperture measures) with glottal aperture as the dependent variable, cuneiform tubercle width as a covariate, phonation mode and phonological voicing as fixed factors, and speaker as a random factor. There was no significant correlation between glottal aperture and cuneiform tubercle width ($p = 0.257$, adjusted $p = 0.149$). Nor was there a main effect of phonation mode: whispered consonants had similar glottal aperture to their spoken counterparts ($p = 0.164$, adjusted $p = 0.199$). There was a main effect of voicing: voiceless consonants, at 32 px, had a 22-pixel (234%) greater glottal aperture than voiced consonants at

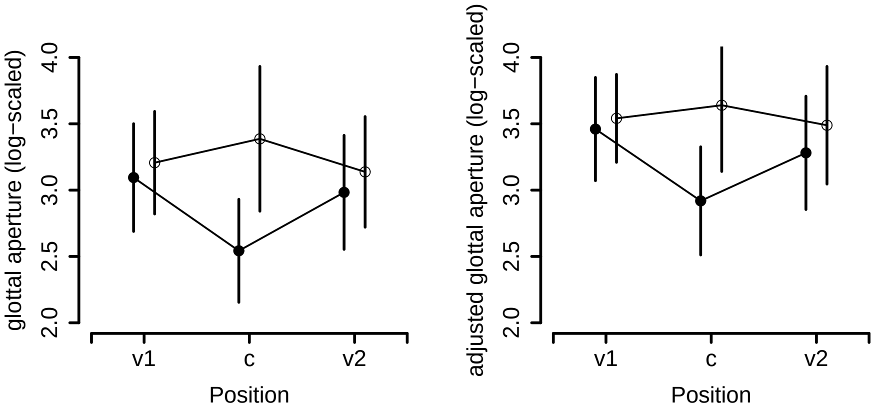


Figure 3.35: Log-scaled glottal aperture plots (left=plain, right=adjusted for cuneiform tubercle obstruction) for whispered obstruents (averaged over all speakers). The points are taken from the vowel preceding the consonant (v1), the consonant itself (c), and the vocalic sequence following the consonant (v2). The filled circles represent voiced tokens, and the open circles represent voiceless tokens. Vertical lines represent one standard deviation above and below each mean.

9 px ($p < 0.001$, adjusted $p = 0.010$)—just as in the analysis for only the spoken consonants.

There were no significant interactions. Crucially, there was no interaction between phonation mode and voicing—the effect of phonological voicing on the consonant's glottal aperture in whisper was not significantly different from the effect of phonological voicing on the consonant in normal speech ($p = 0.098$, obstruction-adjusted $p = 0.401$). Note that, although both the plain and the adjusted glottal aperture measures agree that there is no interaction, the plain measure has a much lower p-value than the adjusted value. Figure 3.36 shows that the adjusted measure has a greater variance than the plain measure for phonologically voiced consonants in whisper. This explains the large difference in p-values, as a greater variance increases the likelihood that a difference between means is due to chance.

The same data is shown log-scaled in figure 3.37. (As the zero values affect the variance of the phonologically voiced consonants in normal speech so much in the log-scaled data, they are excluded from this figure.) Note that the variances are very similar (with the exclusion of the zero values). Note also that the apparent glottal aperture is greater for phonologically voiced tokens in whisper than

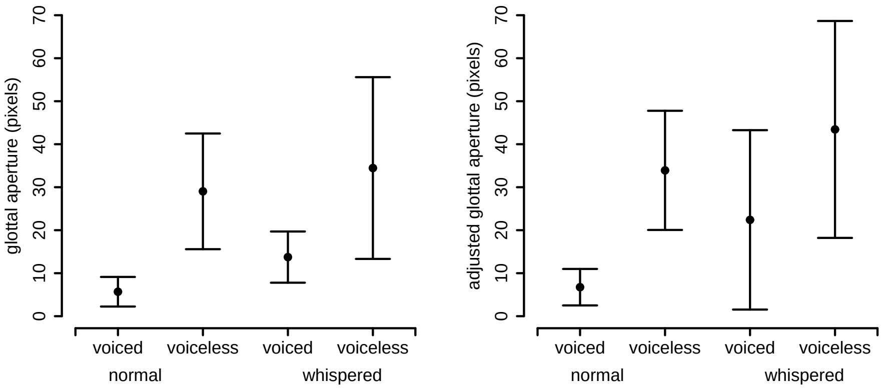


Figure 3.36: Mean glottal aperture in the four crossed conditions: phonologically voiced and voiceless consonants, in normal and whispered speech. Error bars indicate one standard deviation above and below the mean. The left plot gives plain glottal aperture; the right plot gives adjusted glottal aperture.

in normal speech. This does not show up in our statistical results because the linear model takes into account the tendency of whispered tokens to have greater apparent cuneiform tubercle width. In other words, the apparent effect in this figure is because the larynx tends to be closer to the camera in whisper than in normal speech.

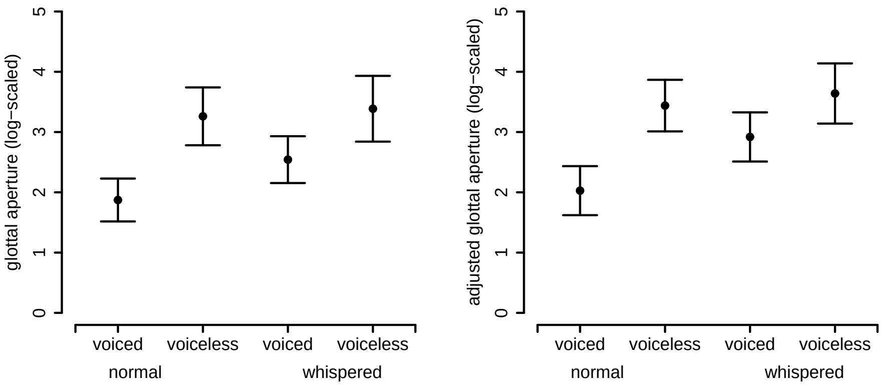


Figure 3.37: Mean log-scaled glottal aperture in the four crossed conditions: phonologically voiced and voiceless consonants, in normal and whispered speech. Error bars indicate one standard deviation above and below the mean. The left plot gives plain glottal aperture; the right plot gives adjusted glottal aperture.

Figure 3.38 illustrates the influence of phonation mode and phonological voicing on glottal aperture, when plotted against the cuneiform tubercle width. Looking at the regression lines, it is clear that phonologically voiceless consonants have greater glottal aperture than phonologically voiced consonants, but the phonation mode (normal or whispered speech) does not affect the glottal aperture of consonants.

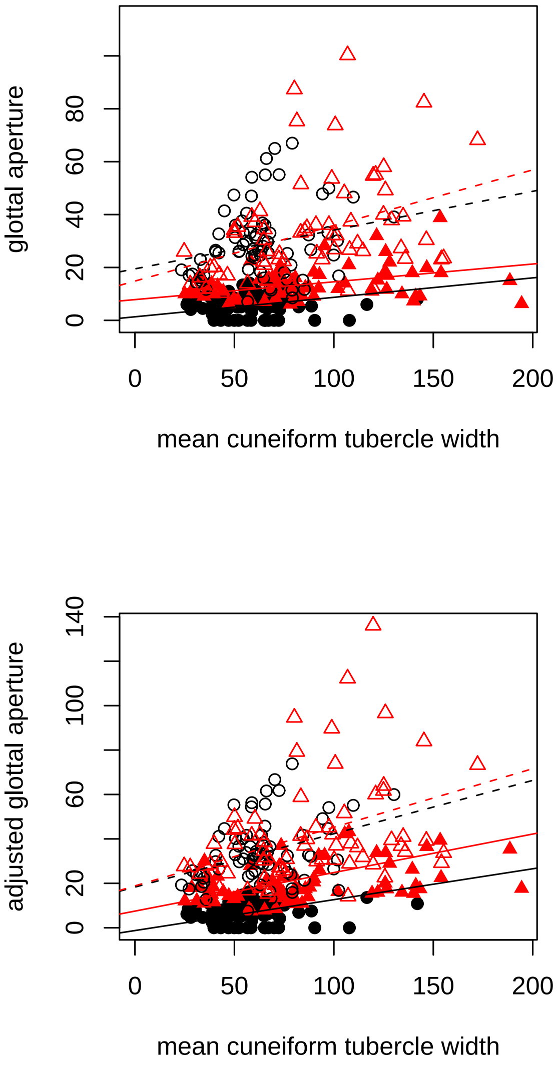


Figure 3.38: Plain (top) and adjusted (bottom) plots of glottal aperture against cuneiform tubercle width for consonants, with regression lines for the separate groups. Spoken items are black; whispered are red. Voiceless consonants are open circles (dashed regression lines); voiced consonants are closed circles (solid regression lines).

The crucial finding here is that there is no significant difference between glottal aperture on consonants in normal speech and in whispered speech. The clear abducted/adducted contrast in normal speech between phonologically voiced segments /p t f s/ and phonologically voiceless segments /b d v z/ is preserved in whispered speech. In fact, not only the contrast but the specific apertures appear to be preserved.

This represents the first quantitative characterization of how glottal aperture corresponds to phonological voicing in whispered speech. Based on previous whisper studies (see section 1.4.1), I expected that the presence or absence of glottal aperture distinctions in whispered phonological voicing contrasts would say something about the acoustic or articulatory basis of the control variable governing glottal aperture. However, such conclusions hinge on the assumption that, in the absence of phonetic voicing, there is no audible consequence of glottal aperture manipulations in whisper. This assumption is tested and refuted in the perceptual study detailed below.

The current results (whatever their meaning for the theory of motor control) verify the expectation of Sweet (1906), Catford (1964), and others for voiceless phonemes (abduction to the normal voiceless glottal state). However, they contradict the speculation about voiced obstruents: while Sweet and others expected speakers to use whisper posture for all phonologically voiced segments, we see speakers clearly trying to give voiced obstruents the same glottal aperture in whisper as they do in normal speech—they use the whispered posture for vowels only, and approximate the voiced posture for phonologically voiced obstruents.

3.7 Perceptual test

The interpretation of the above results depends crucially on whether there are audible cues to phonological voicing generated by the gestural contrast observed. A perceptual test was performed using the acoustic portion of the recordings measured above to determine if such audible cues exist.

3.7.1 Dataset

All productions measured in the above study were collected. However, some had more than two repetitions. In order to maintain a balanced design for the perceptual study, I randomly discarded excess tokens so that there were two repetitions in each cell of the crossed design. That left 2 spoken and 2 whispered repetitions of each of 8 sentences by each of 9 speakers. For two tokens, only one repetition had been recorded. Each of these was included twice to balance the design.

In Mills (2003), a similar test showed better-than-chance perception; but that performance was attributed to durational cues. Voiceless obstruents have significantly greater duration than voiced obstruents. Because duration (a supralaryngeal articulatory cue) would confound the current question (whether laryngeal gestures themselves aid discrimination), the recorded tokens were modified to remove any durational cues to voicing. Tokens illustrating stops (/p b t d/) were truncated to begin just before the release. Tokens illustrating fricatives (/f v s z/) were truncated to include only the final 100 ms of the fricative. If duration is used by listeners to judge consonant voicing, this truncation may bias the answers in favour of the phonologically voiced alternatives. However, this bias cannot produce better-than-chance performance (it will equally bias phonologically voiced and voiceless productions).

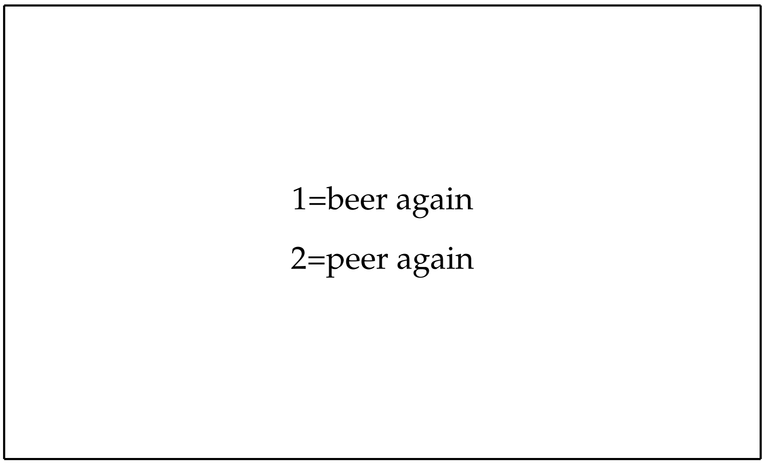
3.7.2 Presentation

The experiment was run using the E-Prime perceptual experiment software (E-Studio 2003). Participants were seated in quiet experimental booths and used high-quality headphones.

The 288 stimuli were presented in random order. A 250 ms pause was included between each response of the listener and the presentation of the next stimulus. Each audio stimulus was accompanied by a visual prompt indicating two options—the word produced (correct), and its voicing minimal pair (incorrect). The visual prompt indicated which key on a keyboard (“1” or “2”) to press for each alternative. There was no means for participants to listen again to an audio stimulus if they were uncertain. Figure 3.39 illustrates a typical visual prompt.

I had previously found that listeners were biased toward a voiceless judgment for whispered tokens Mills (2003); in that study, however, the voiceless option was always presented as option “1”, so the bias may have been for response “1” rather than for perceived voicelessness. Therefore, I randomized the order of the alternatives in the current study—the voiceless alternative was “2” as often as it was “1”.

Three blocks of stimuli were presented to participants. Each block contained the full set of stimuli, in independently randomized order.



1=beer again

2=peer again

Figure 3.39: Sample slide presented to participants.

A full session lasted approximately 30 minutes.

3.7.3 Participants

The listeners were eleven adult native speakers of Canadian and American English (similar to the Canadian English of the speakers from whom the stimuli had been recorded), all naive to the purposes of the study. None of the participants had spent much time in contact with other varieties of English before adulthood.

Participants 3, 7, and 12 had studied phonetics at the graduate level[7]. Participants 10 and 13 had taken an undergraduate course in phonetics. No other participant had any formal phonetic training.

One participant (11) reported that one ear might be less sensitive than the other; others were not aware of any hearing problems. During the sessions of three participants (9, 10, and 11), building work involving jackhammers intruded for a substantial portion of the session.

[7] Some participant numbers are higher than the number of participants because some individuals withdrew from the study after being assigned a number.

These potential confounds of intrusive noise and minor hearing loss would tend to produce results closer to random behaviour (favouring H_0). Data from participants 9, 10, and 11 are therefore checked to see if they diverge from the overall tendency.

The phonetic training of listeners 3, 7, 10, 12, and 13 might be expected to artificially boost their perceptual performance, leading us to reject the null hypothesis when perhaps we should not. Their data are also checked against the overall results to see if they diverge.

Participants were instructed to identify which sentence fragment they thought they heard, and to pick randomly if they were completely uncertain.

Most participants reported feeling that several tokens had the initial consonant completely removed (labial stops especially); several also reported that for some tokens they thought they heard a word that was not among the options (“gear” instead of “beer”, for example).

3.7.4 Results and analysis

Overall, listeners correctly identified tokens in normal speech 93% of the time (n=4752, χ^2=3441.49, df=1, p<0.001). In whispered speech, they also performed significantly above chance, with correct identifications 65% of the time (n=4752, χ^2=452.26, df=1, p<0.001). The difference between listener performance in normal and whispered speech is significant (n=7507, χ^2=221.33, df=1, p<0.001). Figure 3.40 illustrates overall performance (proportion correct responses) in this perception task.

The five listeners with phonetic training performed similarly. They scored 93% in normal speech (n=2160, χ^2=1574.23, df=1, p<0.001) and 67% in whispered speech (n=2160, χ^2=236.02, df=1, p<0.001). The difference between performance on normal and whispered speech is significant (n=3439, χ^2=92.82, df=1, p<0.001).

The three listeners exposed to significant background noise also performed similarly. They scored 91% in normal speech (n=1296, χ^2=860.44, df=1, p<0.001) and 65% in whispered speech (n=1296, χ^2=111.42, df=1, p<0.001). The difference between performance on normal and whispered speech is significant (n=2014, χ^2=56.72, df=1, p<0.001).

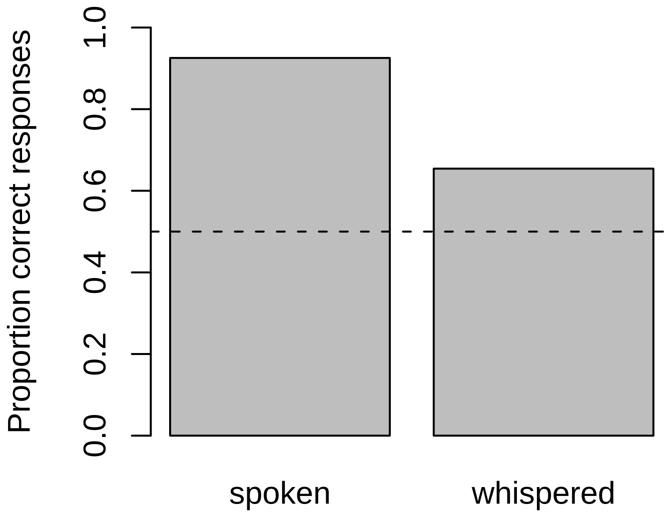


Figure 3.40: Overall perception scores on endoscopic recordings. The horizontal dotted line corresponds to chance performance (50% correct).

Subdividing the data by manner of articulation, as depicted in figure 3.41, an interesting pattern emerges. In normal speech, stops were identified with 99% accuracy (n=2376, χ^2=2257.52, df=1, p<0.001), but fricatives with only 86% accuracy (n=2376, χ^2=1256.73, df=1, p<0.001)—a difference that is significant (n=4398, χ^2=19.65, df=1, p<0.001). In whispered speech, stops were identified with 75% accuracy (n=2376, χ^2=584.04, df=1, p<0.001) but fricatives were identified with only 56% accuracy (n=2376, χ^2=34.91, df=1, p<0.001)—also a significant difference (n=3109, χ^2=63.69, df=1, p<0.001).

Alveolar stops /t/ and /d/ are discriminated slightly better in whisper than labial stops /p/ and /b/—scores are 81% for alveolar stops and 72% for labial stops. This is consistent with listeners' subjective impressions that, for the /p b/ tokens, the initial consonant sometimes seemed to be missing entirely. Acoustically, this may be due to the fact that the formant transitions for labial consonants are less pronounced than those for alveolar consonants; thus, in the absence of any other acoustic cue, the labial plosives sometimes seem to disappear entirely.

No systematic difference in perception is observed between the labial and alveolar fricatives.

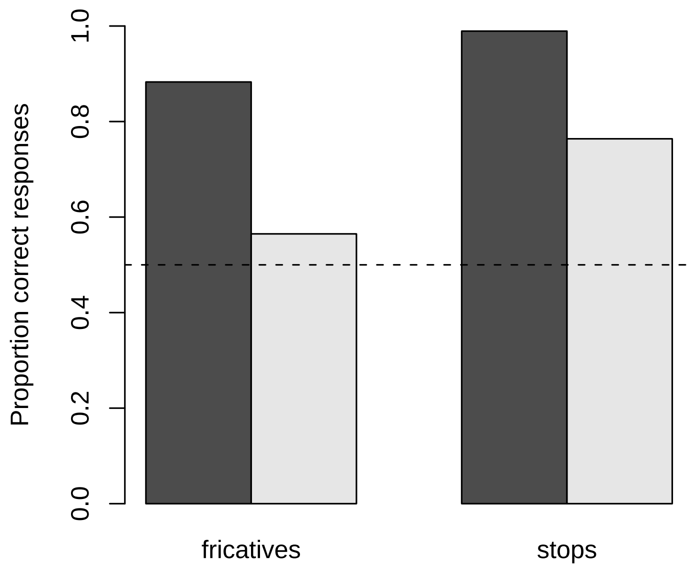


Figure 3.41: Perception scores on stop and fricative contrasts in normal and whispered speech. The horizontal dotted line corresponds to chance performance (50% correct). The dark bar indicates normal tokens; the light bar indicates whispered tokens.

The main trends to note in the above data are that (a) listeners are better at distinguishing voicing pairs in normal speech than in whispered speech; (b) even without duration as a cue, listeners perform better than chance at distinguishing all labial and alveolar stop and fricative voicing pairs; and (c) perceptual performance on fricative pairs suffers more from the removal of phonetic voicing and duration cues than does performance on stop pairs.

3.8 Discussion

This chapter presents the results of an investigation into the nature of glottal aperture control in whispered speech. A detailed quantitative methodology is introduced for extracting useful numerical values from complex, multivariate video data that contains several distortions and potential confounds to measurement.

The data yield a characterization of glottal aperture patterns as they relate to phonological voicing contrasts in Canadian English in both normal and whispered speech. Glottal aperture distinctions are seen between phonologically voiced and voiceless obstruents in both normal and whispered speech; moreover, the distinctions are the same in whispered speech as they are in normal speech, contrary to the expectations of several previous researchers. Relating this back to the diagrammed hypotheses in figure 3.1, our results fit neither of the proposed hypotheses. Figure 3.42 presents the observed outcome (alongside the pattern for normal speech).

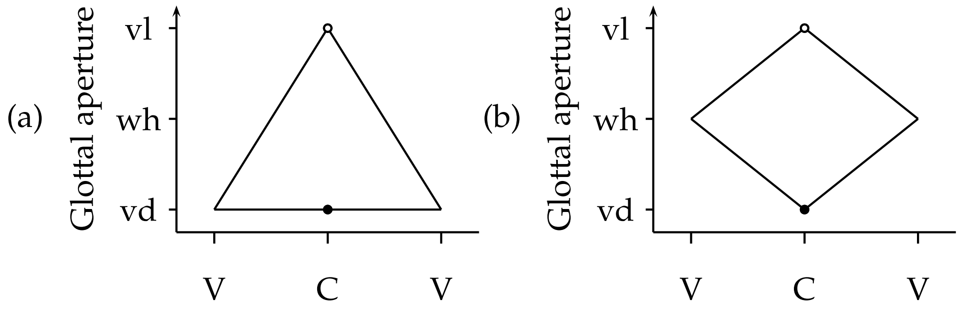


Figure 3.42: Diagrammatic representation of glottal aperture in voicing contrasts: (a) normal speech; (b) whispered speech—abduction and adduction gestures present. Open circles represent phonologically voiceless obstruents; closed circles represent phonologically voiced obstruents.

A perceptual test demonstrates that, even in the absence of phonetic voicing, the glottal aperture distinctions seem to generate audible differences between the phonologically voiced and voiceless obstruents.

We therefore cannot conclude that the glottal distinctions observed are due to an articulatory specification for the motor control variable governing the larynx. The control variable *may* have an articulatory specification. However, it may also have an acoustic specification which exploits the remaining acoustic cues that the listeners are picking up on.

Further discussion is offered in chapter 6.

CHAPTER 4

Prosodic measures

Chapter 5 presents an experiment in which a prosodic feature is manipulated, and associated acoustic properties are observed. The current chapter gives the details of how those acoustic properties are measured.

Five acoustic parameters have been implicated in the communication of relative prominence in English: f_0 and duration (Fry 1955, 1958, Cooper, Eady & Mueller 1985, Eady & Cooper 1986, Aylett & Turk 2006), amplitude (Traunmüller & Eriksson 2000), spectral tilt (Sluijter & van Heuven 1996, Heldner 2001, Remijsen 2001, Wouters & Macon 2002), and vowel quality (Aylett & Turk 2006). All five were measured in this study. The following sections present the details of how each was measured.

4.1 Duration

Duration measures taken include the duration of the entire target word as well as the duration of the accented syllable. In order to confirm that varying speech rate does not confound our results, an easily-segmented sequence preceding the target word was also marked. This control sequence was designed to contain a prenuclear pitch accent (generally, on the content word previous to the target word). These control duration sequences are indicated in table 4.1.

In order to obtain clear segmentations, only the first syllable of the control words "golfing" (for target word "father") and "searching" (for target word "marshes") were used.

Sample sentence	Target word	Control
Their seatbelts will be fastened.	fastened	seatbelts
She can cycle fastest.	fastest	cycle
They're golfing with their father.	father	golf(ing)
They're searching the marshes.	marshes	search(ing)
It's on in the morning.	morning	on in the
He checked his main sources.	sources	check[ed]
They're going surfing.	surfing	going
They ordered some sushi.	sushi	order[ed]
She's writing a thesis.	thesis	writing
He's very thirsty.	thirsty	very

Table 4.1: Sections of sentences used as control durations.

Condition	Control		
Statement	He	checked	his main sources.
Question	Did he	check	his main sources?
Statement	They	ordered	some sushi.
Question	Did they	order	some sushi?

Table 4.2: Control words varied across conditions for two of the target words.

Tokens for "morning" were not ultimately used, because of the lack of a consistent prenuclear accent mentioned in section 4.2 above.

For target words "sources" and "sushi", the control word varied between inflected for statements and uninflected for questions, as illustrated in table 4.2. Because the key statistical analysis is within sentence types, these differences should not confound our results.

Durations of segments were measured on the basis of acoustic landmarks common to both normal and whispered speech. The onset and offset of voicing could not be used as landmarks to identify segment boundaries, because any measurements derived from phonetic voicing boundaries would not be comparable between normal and whispered speech.

The main landmarks we used are vowel formant onset and offset. Most target words contain mainly sequences of alternating vowels and voiceless fricatives. The onset and offset of vowel formants are used (rather than voicing) for these boundaries (figure 4.1).

Note that, as in figure 4.1, some tokens contained plosives. In order to facilitate the formant analysis, boundaries were taken at the end of the release burst,

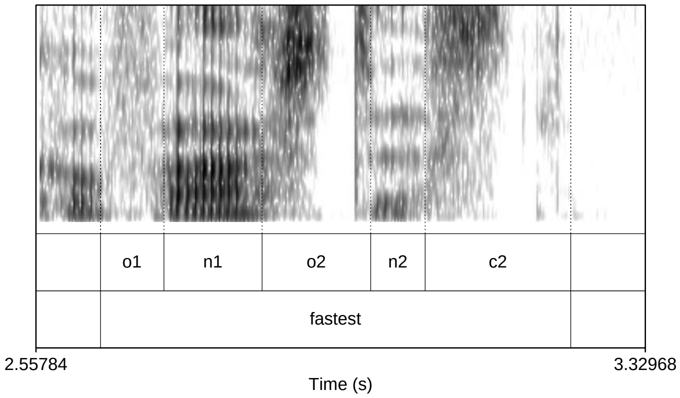


Figure 4.1: Segmented token of "fastest" spoken by speaker 4. Labels for individual segments indicate syllable position (o=onset, n=nucleus, c=coda) and the syllable that the segment is in (1 or 2). The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

rather than at the beginning (as in the endoscopy study). Release bursts contain high frequency noise which would disrupt formant tracking and possibly skew formant tracks.

The same criteria were used in segmenting whispered tokens (figure 4.2).

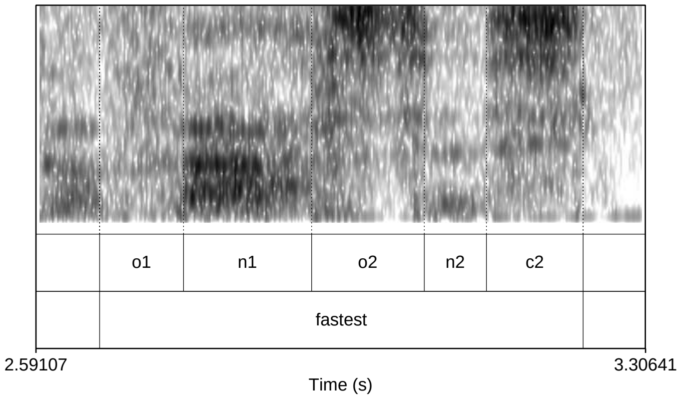


Figure 4.2: Segmented token of "fastest" whispered by speaker 4. Labels for individual segments indicate syllable position (o=onset, n=nucleus, c=coda) and the syllable that the segment is in (1 or 2). The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

Some words contain nasals bordering vowels; again, vowel formant onset and offset were used (see figures 4.3 and 4.4).

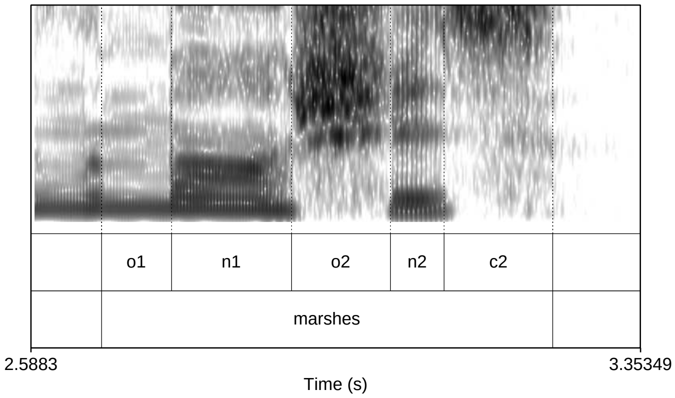


Figure 4.3: Segmented token of "marshes" spoken by speaker 5. Labels for individual segments indicate syllable position (o=onset, n=nucleus, c=coda) and the syllable that the segment is in (1 or 2). Note the boundaries on either side of segment "o1" ([m]). The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

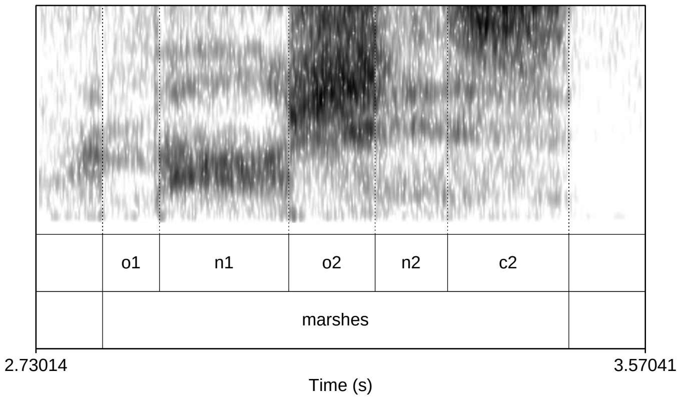


Figure 4.4: Segmented token of "marshes" whispered by speaker 5. Labels for individual segments indicate syllable position (o=onset, n=nucleus, c=coda) and the syllable that the segment is in (1 or 2). Note the boundaries on either side of segment "o1" ([m]). The vertical (frequency) axis goes from 0 Hz to 5000 Hz.

4.2 Fundamental frequency

The next parameter to describe is the magnitude of the pitch accent on an accented syllable.

Prosodic prominence is generally signalled by a high f_0 peak *relative* to surrounding prosodic constituents (Portele & Heuft 1997, Terken & Hermes 2000), rather than by absolute f_0.[1]. I therefore measured not only the f_0 peak of the accented target word (nuclear pitch accent), but also the f_0 peak of the word with the previous pitch accent in each utterance (prenuclear pitch accent) and the intervening f_0 minimum.

For every token, glottal pulses were marked by first running an automatic tracking algorithm, and then hand-checking and adjusting bad tracks. Figure 4.5 illustrates the locations of the pulse marks on the accented syllable of "fastened".

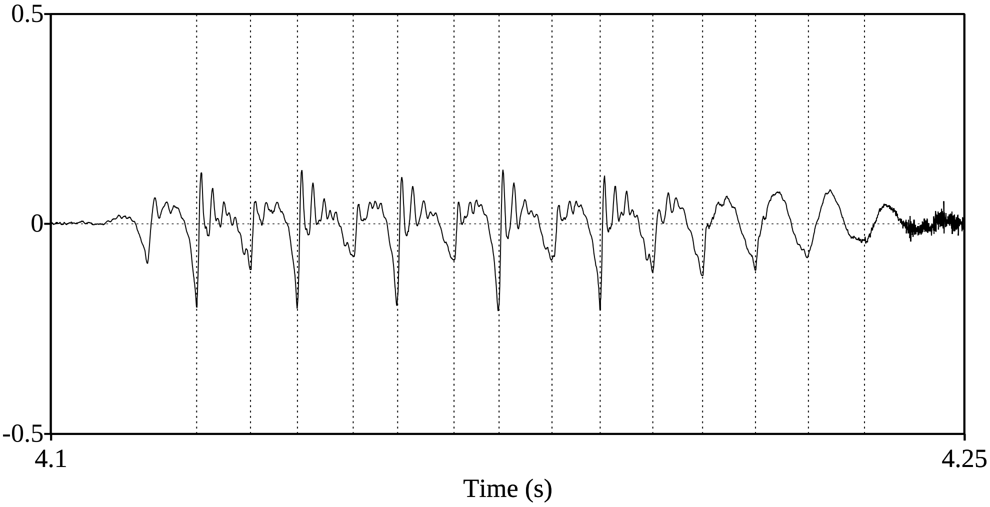


Figure 4.5: Pulse marks on the accented syllable of "fastened". The solid trace is the waveform; the horizontal dashed line is the reference pressure; the vertical dashed lines indicate the times at which glottal pulses are marked.

Ideally, we would use a smoothed contour that abstracts away from segmental perturbations of f_0 and extrapolates over unvoiced sections of the signal. The MOMEL algorithm (Hirst & Espesser 1993) provides smoothing of segmental perturbations, as well as extrapolation across voiceless sections of the signal,

[1] Following preliminary results from Gussenhoven & Rietveld (1988), Ladd, Verhoeven & Jacobs (1994) found that the perceived prominence of an accent is sometimes positively, and sometimes negatively, correlated with the relative f_0 height of a preceding accent. Accents with high f_0 have negative correlations (perceived as more prominent when the f_0 of the preceding accent is lower); those with low f_0 have positive correlations (perceived as more prominent when the f_0 of the preceding accent is higher). These studies do not, however, look at whether this perceptual inversion is reflected in speakers' productions. That is, we do not know whether speakers manipulate prominence sometimes by varying f_0 directly with the preceding peak, and sometimes by varying f_0 inversely with the preceding peak. For the current study, I use two separate measures of relative prominence: the nuclear peak relative to the prenuclear peak, and the nuclear peak relative to the prenuclear valley.

based on fitting quadratic spline segments to the data. I used an implementation of this algorithm for Praat to identify maxima and minima in the f_0 contour[2]. Figure 4.6 illustrates the output of this algorithm (smooth line), plotted alongside the f_0 values derived from adjascent pairs of glottal pulse marks (individual dots). This algorithm yielded mixed results: for some tokens (such as the one illustrated), both prenuclear and nuclear accent peaks were tracked well. For many tokens, the tracking algorithm was unable to generate a contour that looked reasonably close to the actual data. Another frequent problem was that the inter-accent valley was generally not marked very well. Note that this token has a falling boundary tone, even though it is a question. This was a common feature of our data—speakers produced a high pitch accent followed by a low, with no rise at the end of the sentence. See section 5.6.12 for results, which show

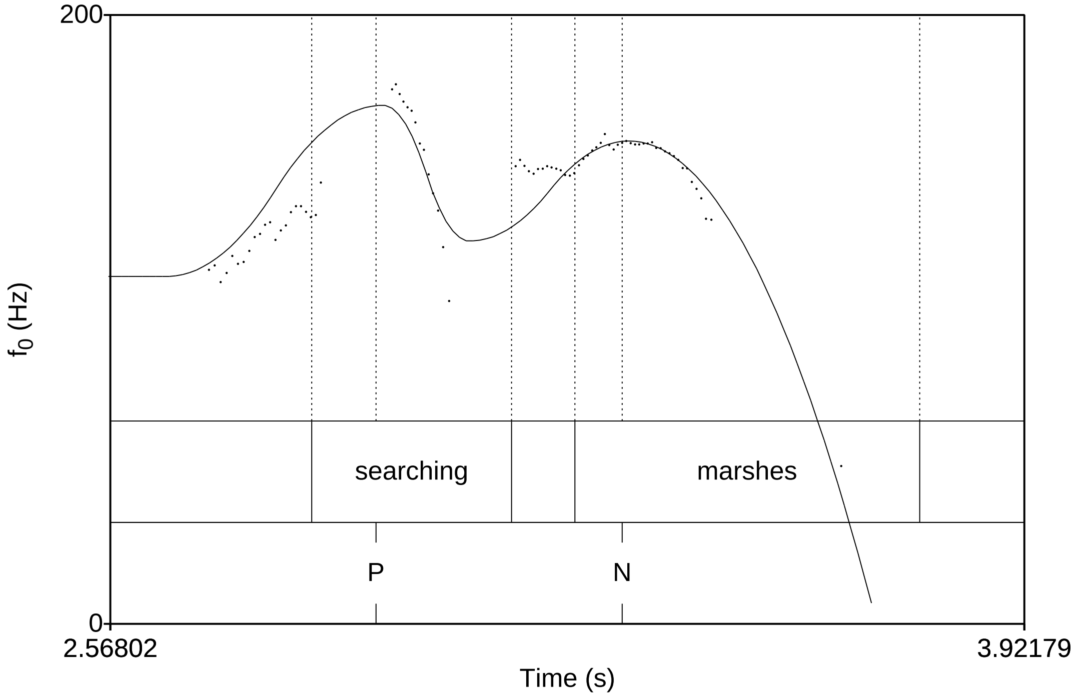


Figure 4.6: Example of f_0 smoothing output for sentence "Are they searching the marshes?" The un-smoothed f_0 contour after manual correction of the voicing pulses is indicated by points (each point represents the period between two glottal pulses). The output of MOMEL algorithm is shown as a curve. "P" indicates the prenuclear accent peak and "N" indicates the nuclear accent peak.

[2] The implementation we used was written by Guillaume Roland, with modifications by Bert Remijsen and further modifications by the current author. The modifications dealt with file structures, and did not affect the actual f_0 smoothing and spline-fitting algorithm.

that speakers did, nevertheless, use f_0 to distinguish statements from questions, at least as far as the pitch accent on the utterance-final word is concerned.

These shortcomings meant that a significant portion of the data did not yield reliable f_0 measurements after MOMEL smoothing. Because of this, a measurement procedure based on the raw f_0 track was followed instead.

From the manually-corrected pulse marks, a series of f_0 values were calculated and plotted. Each point in the plot represents the frequency calculated from the period between two pulses. Examples of the f_0 plot from this method are seen in figures 4.7, 4.9, and 4.8.

For tokens where a clear, smooth peak-valley-peak pattern was evident in the f_0 track, the peaks and valley were taken as the frequency points at the local maxima and minimum, as shown in figure 4.7. Note that, in order to reduce confounds from segmental f_0 perturbations, an f_0 point was only taken as a true maximum or minimum if it was part of a sequence of at least two simlar f_0 points. This avoids undue influence from obvious outliers, three of which can be seen in figure 4.7 (not marked—there is one outlier adjascent to each of the labelled turning points).

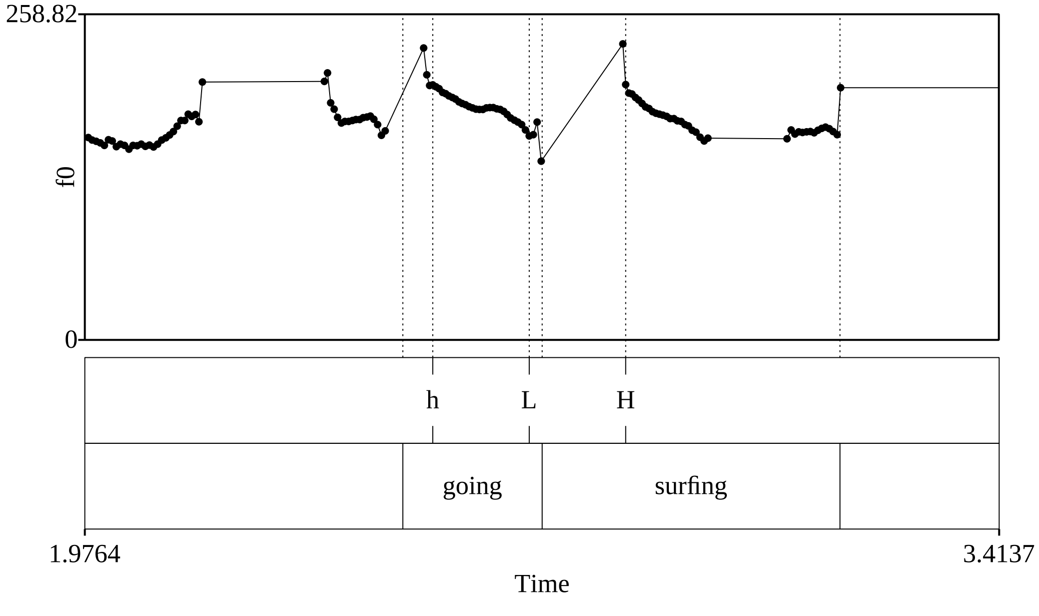


Figure 4.7: Token with marked-up f_0 contour. The pitch points are marked "h" (prenuclear accent f_0 peak), "L" (inter-accent f_0 minimum), and "H" (nuclear accent f_0 peak).

The words of interest in this study are utterance-final, so there is a large amount of creakiness. As a result, non-smooth f_0 tracks were common. Specifically, an apparent diplophonia was often observed, as in the first syllable of "fastened" in figure 4.8. In such cases, I treated the upper of the two alternating f_0 tracks as a valid source of peak values. For example, in this token, the nuclear accent peak is taken as the highest point in the upper part of this alternation.

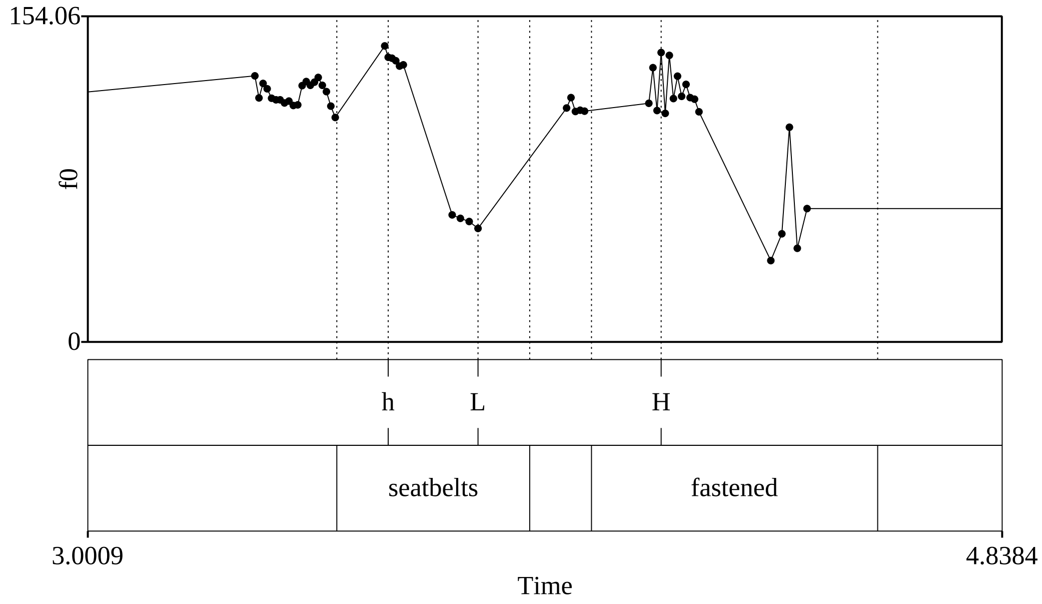


Figure 4.8: Token with diplophonic f_0 contour.

In some cases, creakiness was so severe that no clear peak or valley could be identified. In figure 4.9, for example, no prenuclear accent is apparent. Also, the obvious low-point is about half the frequency of the adjascent portions of speech, with an abrupt rather than gradual transition. This suggests a qualitative rather than quantitative shift in phonation, rendering the use of this as a "valley" less valid. Tokens like this were discarded.

The sentences for the target word "morning"—"It's on in the morning" and "Is it on in the morning?"—did not have a consistent prenuclear accent on the word "on" as expected. Without a prenuclear accent to compare the nuclear accent to, I had to exclude all "morning" tokens from the analysis.

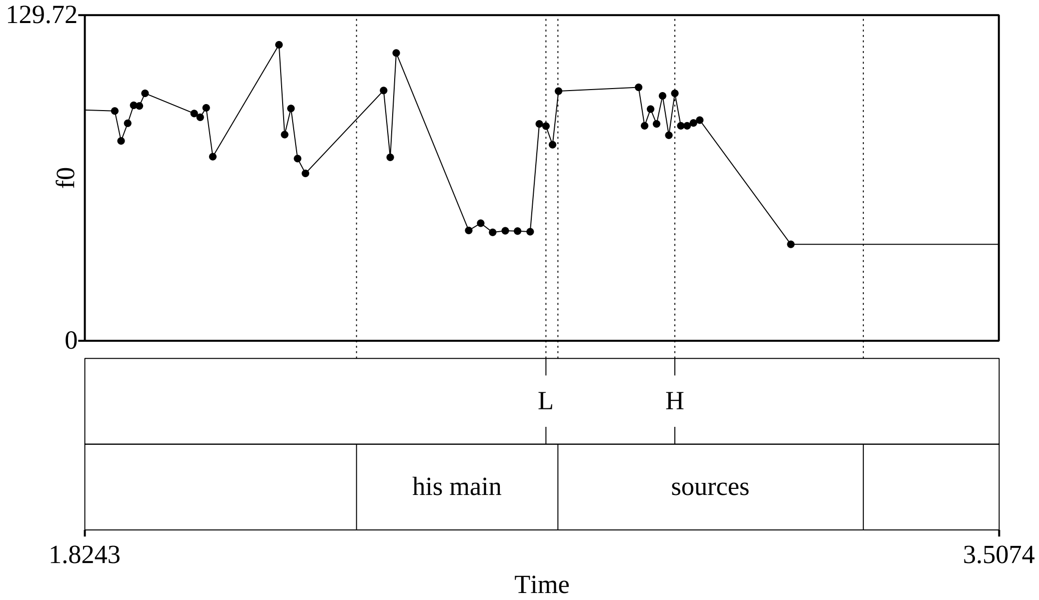


Figure 4.9: Token with unusable f_0 contour.

Frequency tracks were produced for spoken statements and questions, but not for whispered statements. Figures 4.10, 4.11, and 4.12 illustrate the overall distribution of frequency measures for the nuclear peak, the prenuclear peak, and the interpeak minimum (respectively). All three show a distinct positive skew.

A logarithmic transform of these data—converting Hertz to semitones—removes this skew. This is shown in figures 4.13, 4.14, and 4.15.

The formula used in this work to derive a logarithmic frequency measure in semitones (f_{log}) from a linear one in Hertz (f_{lin}) is given in 4.1. This formula uses a reference frequency of 100 Hz. Frequency values yielded therefore represent the number of semitones above 100 Hz. A semitone is one twelfth of an octave—twelve semitones represents a doubling of the linear frequency.

$$f_{log} = \frac{12 \ln \left(\frac{f_{lin}}{100} \right)}{\ln 2} \tag{4.1}$$

Analyses in chapter 5 are performed on the log-scaled measurements.

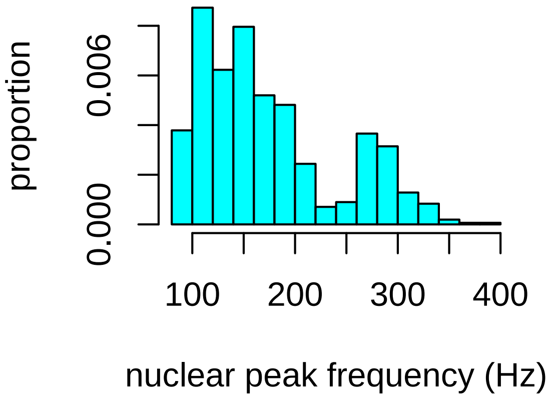


Figure 4.10: Distribution of nuclear peak frequencies.

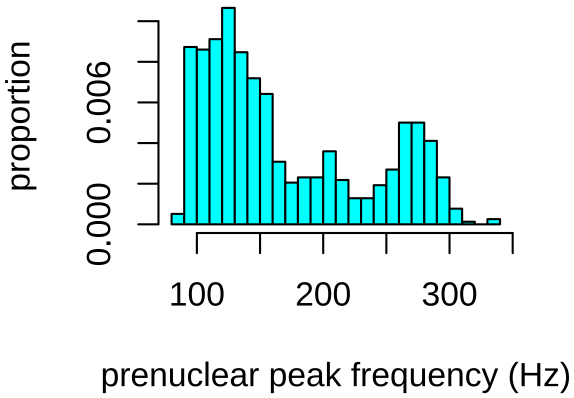


Figure 4.11: Distribution of prenuclear peak frequencies.

4.3 Amplitude

The third acoustic measure relevant to contrastive accent is the amplitude (Traunmüller & Eriksson 2000).

For this data, we take amplitude as the maximum excursion from zero pressure over the duration of the vowel, as illustrated in figures 4.16 and 4.17.

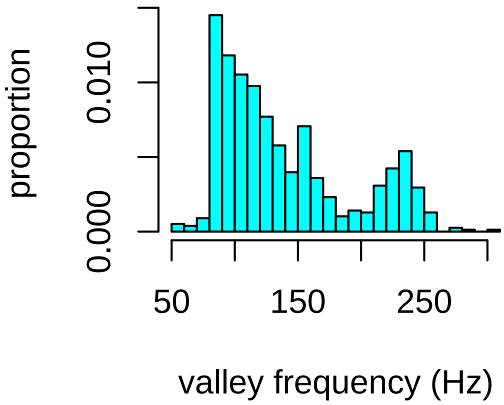


Figure 4.12: Distribution of interpeak minimum frequencies.

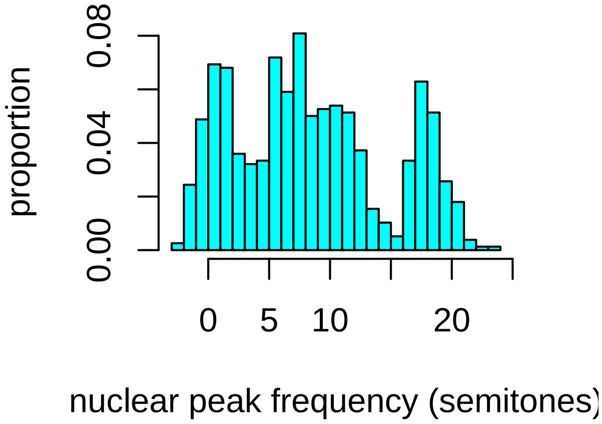


Figure 4.13: Distribution of log-scaled nuclear peak frequencies.

Amplitude measurements in decibels (L_P) were derived from the peak pressure measures (P) using the standard formula (equation 4.2). In the absence of an absolute standard to use as a reference pressure (P_0), I used the maximum excursion from zero pressure in the vowel with prenuclear accent (in the control

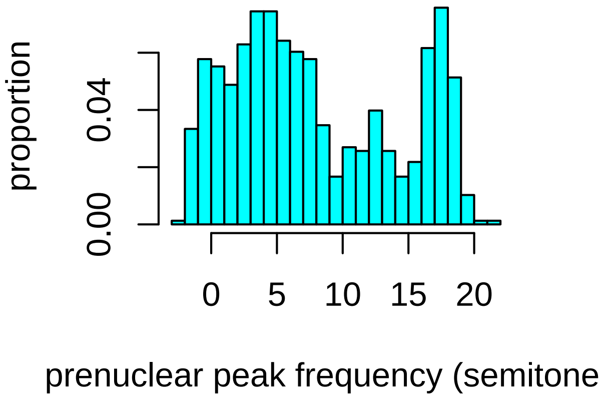


Figure 4.14: Distribution of log-scaled prenuclear peak frequencies.

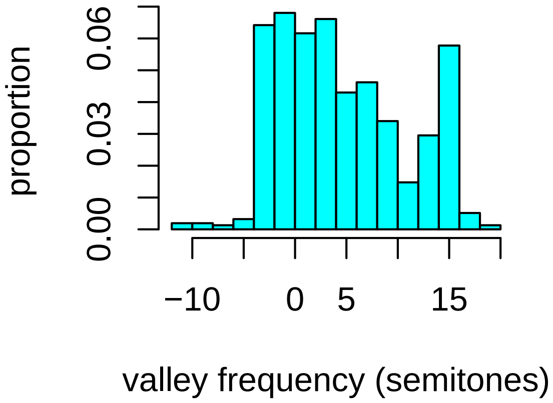


Figure 4.15: Distribution of log-scaled interpeak minimum frequencies.

words identified in table 4.1).[3]

$$L_P = 20 \log_{10} \left(\frac{P}{P_0} \right) \tag{4.2}$$

[3] I used Praat's `Get absolute extremum...` function (Boersma & Weenink 2005), with the standard `Sinc70` interpolation (fitting a sinc function to the waveform using 70 samples each side of the sample with the greatest absolute value; see Praat online help for full details).

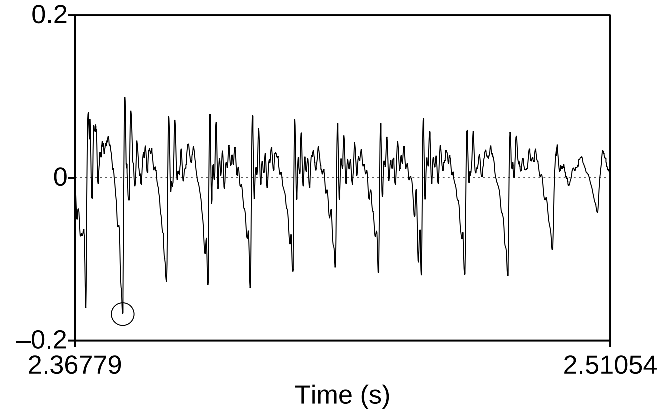


Figure 4.16: Example measurement of amplitude on a spoken vowel. The absolute extremum is indicated with a circle.

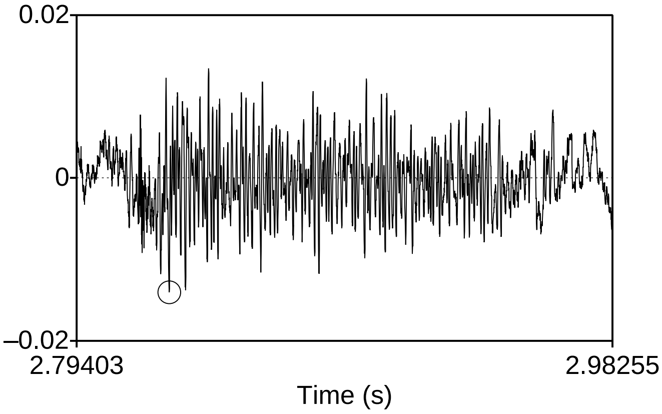


Figure 4.17: Example measurement of amplitude on a whispered vowel. The absolute extremum is indicated with a circle.

Because each token's amplitude measure is normalized against a value within the same sentence, varying recording levels between speakers will not affect this measure.

4.4 Vowel quality

The automated formant-tracking algorithm used in the Praat software (Boersma & Weenink 2005) was used to track the first three formants of each accented vowel. Visual inspection was used to verify the algorithm's accuracy. Gross errors such as treating two closely-spaced formants as one were corrected by adjusting the tracker's parameters on individual tokens until the track fell within the visually-apparent range of each formant throughout the vowel. The average of each track over 50 ms at the midpoint of the vowel (or as close as possible to the midpoint, if an audio tick disrupted the track) was taken as the target formant value for that syllable.

Vowel quality is expected to become less centralized in contrastive (more prominent) contexts (Aylett & Turk 2006). This means that we expect a different direction of effect on the formants depending on where in the formant space a vowel normally lies.

I began by comparing the mean F1 and F2 values of each vowel in the data to the overall formant means, as a proxy for a completely centralized vowel (see figure 4.18).[4] Table 4.3 shows the direction of difference of F1 and F2 for each

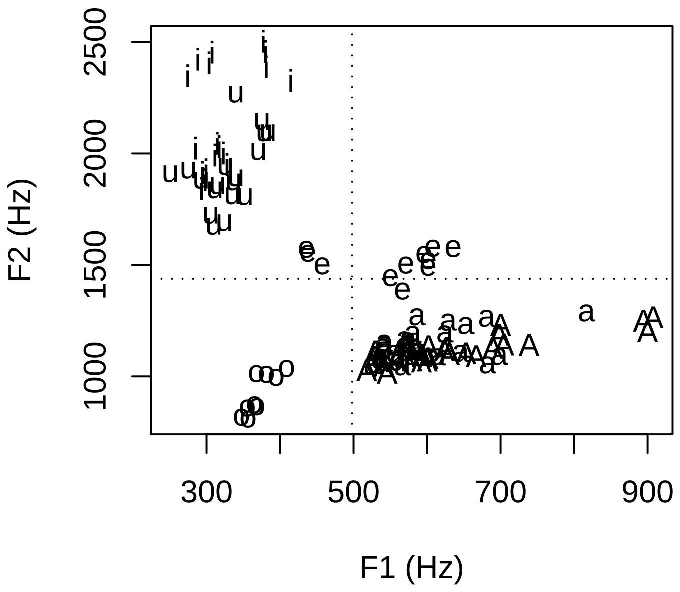


Figure 4.18: Vowels plotted by formant values. The dotted lines represent the average F1 and F2 values of all data points. ASCII approximations are used for the vowels: "i" for [i], "u" for [u], "e" for [ə], "o" for [ɔ], "a" for [æ], "A" for [ɑ].

vowel represented in the accented syllables of the current data. These means are taken from the statements in normal speech with non-contrastive contexts. (Note

[4] An alternative means of defining the centralized vowel would have been to take the actual measurements of the accented vowel in, for example, "thirsty" /θəːsti/; results would have been similar.

	F1	F2	Words
/æ/	above	below	fastened, fastest
/ɑː/	above	below	father, marshes
/əː/	above	above	surfing, thirsty
/i/	below	above	thesis
/ɔː/	below	below	morning, sources
/u/	below	above	sushi

Table 4.3: Difference of vowel formants from mean (all differences have $p < 0.05$ in individual single-sample t-tests).

that some vowels are represented by more than one target word—such as /æ/ in both "fastened" and "fastest".) The table shows /æ/, /ɑː/, and /əː/ with F1 higher than the overall mean and /i/, /ɔː/, and /u/ with F1 lower than the overall mean. It also shows /əː/, /i/, and /u/ with F2 higher than the overall mean and /æ/, /ɑː/, and /ɔː/ with F2 lower than the overall mean. I left schwa (/ə/) out of further analysis because it is generally considered to be reduced, despite the fact that in the current dataset it departs significantly from the overall mean (having higher F1 and F2).

The measures of F1 and F2 frequency that were used in the spectral tilt calculations also served as indicators of vowel quality.

4.5 Spectral tilt

Spectral tilt is the overall distribution of energy in the spectrum; it is also described as the rate at which intensity falls off as frequency increases.

Studies have shown spectral tilt to be involved in the perception and production of lexical stress (Sluijter & van Heuven 1996, Remijsen 2001). Differences between modal and turbulent (breathy, whispered) phonation modes are also reflected in spectral tilt: turbulent excitation produces proportionally more high-frequency energy than modal excitation does (Hanson 1997, Hanson & Chuang 1999, Mills 2003).

Obviously, formant locations also affect the distribution of energy in the spectrum. A high second formant, for example, generates a spectrum with more energy in the high frequencies than there is in a vowel with a low second formant.

I will not deal here with factors external to the speaker. Telephone transmission generally cuts out low-frequency energy; muffling speech with a hand or duct tape dramatically attenuates the high-frequency portions of the signal.

Several different approaches to measuring spectral tilt have been explored in the literature.

4.5.1 Energy bands

Sluijter & van Heuven (1996) use the simple strategy of identifying four frequency bands, and measuring the average energy in each band. They then perform a separate univariate analysis on each band and compare the results to see which part of the spectrum is most affected by lexical stress in Dutch. A similar approach is taken by Remijsen (2001) in Ma'ya, an Austronesian language, and by Mills (2003), looking at Scottish English voicing contrasts in whispered speech.

This method has the advantage of being simple to perform. However, in order to avoid the confound of varying formant frequencies, only a single vowel quality can be used and the frequency bands must be set so that no formant is too close to the boundary between two bands.

For the current study, I would prefer to obtain a value representing the source spectrum, independent of the particular vowel being produced, so that we can directly compare the spectral tilt of an [i] with that of an [o] without worrying about the confound of shifting formant frequencies. The technique of Sluijter & van Heuven is not suitable for cross-vowel comparisons.

It is also numerically inelegant, in that it yields multiple values rather than a single "spectral tilt" value.

4.5.2 Regression lines

In one study (Kochanski, Grabe, Coleman & Rosner 2005), the authors fit a regression line to the spectrum, and use the slope of the line to represent spectral tilt. In order to have a measure with more perceptual relevance, they rescale both frequency and power in the spectrum before doing regression. They rescale the frequency to Barks, and power to the cube-root of power ($\sqrt[3]{\text{power}}$). They generate a best-fit line across (Bark-scaled) energy bins from 500 Hz to 3000 Hz.

While the values yielded by this method is a "slope" value, it is not clear (for our purposes) that it represents a property of the source. Variations in formant values can be expected to significantly affect this slope independently of any changes in the source spectrum.

4.5.3 Voicing-based measures

Some measures are directly or indirectly based on the presence of voicing.

Campbell & Beckman (1997) measure spectral tilt as the "intensity ratio between the first and second harmonics (H2-H1 in dB)". This measure would be very sensitive to F1 movement (low F1 would increase the ratio), thus failing to separate source and filter.

Two studies have implemented a measure based on comparing the energy at f_0 to the overall energy of the spectrum (Traunmüller & Eriksson 2000, Heldner 2001). They produce a separate copy of the signal which is dynamically low-pass filtered at $1.5 \times f_0$. Their spectral emphasis measure is calculated as SPL$_{all}$ − SPL$_f 0$.

This has the advantage of yielding a single value to represent the proportion of energy in the signal that is low-energy. It is also partly insensitive to variations in formant frequencies, except that an F1 proximate to f_0 would probably increase the amplitude of the f_0 energy band.

Like the other measures above, this one does not specifically isolate source spectrum properties. Its value is an amalgam of source and filter properties.

4.5.4 Acoustic models

Ultimately, what is needed for our purposes is a measure of spectral tilt that can systematically eliminate the influences of the supraglottal filter on the spectrum, leaving just the source properties to be compared.

Two research lines have pursued this approach. One, followed by Hanson (1997), Hanson & Chuang (1999), compares the amplitudes of the first harmonic ($H1$) and the third formant ($A3$). First, however, formulas are applied to correct for the influences of the nearby formants:

$$
\begin{aligned}
H1* &= H1 - 20\log_{10}\left(\frac{F1^2}{F1^2 - f^2}\right) && (4.3)\\
A3* &= A3 + 20\log_{10}\left(\frac{[1-(F3/F1)^2][1-(F3/F2)^2]}{[1-(F3/F1_{\text{ref}})^2][1-(F3/F2_{\text{ref}})^2]}\right) && (4.4)
\end{aligned}
$$

Where *f* is the frequency of $H1$; $F1$, $F2$, and $F3$ are the frequencies of the first three formants; and $F1_{\text{ref}}$ and $F2_{\text{ref}}$ are the frequencies of reference formants—the formants $F1$ and $F2$ in a tube the length of the speaker's vocal tract, but of completely uniform diameter.

The main disadvantage of Hanson's approach for our study is that it relies on a measurable H1 (first harmonic peak), which is not present in whispered speech.

An alternative method which uses a more complete and explicit acoustic model than that of Hanson is that introduced by Fulop, Kari & Ladefoged (1998), also used by Guion, Post & Payne (2004). They use the work of Fant (1960) to mathematically model all of the factors contributing to the speech spectrum.

This model identifies six main contributors to the observed spectrum of speech:

1. The first formant
2. The second formant
3. The third formant
4. The remaining formants
5. The source spectrum
6. The radiation of the sound from the mouth

While in principle all formants could be independently measured, it is only the first three that vary enough in speech to be worth measuring individually. Equation 4.5 gives the function for the contribution of a formant of frequency F and bandwidth b to the overall spectrum (amplitude in dB as a function of frequency (f) in Hz).

$$
\text{dB}(f) = 20\log_{10}\frac{F^2 + (b/2)^2}{\sqrt{(f-F)^2 + (b/2)^2} \times \sqrt{(f+F)^2 + (b/2)^2}} \qquad (4.5)
$$

The total contribution of the remaining formants is estimated by the "catch-all" formula 4.6. This formula increases with the fourth power of f (very fast). This renders it unrealistic for very large f, but it is an adequate approximation within the range used for speech analysis.

$$\text{dB}(f) = 0.72(f/492)^2 + 0.0033(f/492)^4 \tag{4.6}$$

The radiation characteristics are assumed to be fixed—the experimental setting is the same across recordings, right down to the distance from the speaker's mouth to the microphone (it is a headset). Fulop et al., following Fant, collapse the radiation and the source formulae into a single equation (4.7).

$$\text{dB}(f) = g\left(20\log_{10}\left(2\frac{f/100}{1+(f/100)^2}\right)\right) \tag{4.7}$$

The parameter g determines the overall slope of this component of the spectrum. Throughout the current work, g is set to 1.0, for a spectral slope of about -6 dB/octave. Acoustic models (Ní Chasaide & Gobl 1997) and empirical investigations (Stevens 1998, p 69) identify this as a reasonable median value for the combined effects of the glottal source spectrum (-12 dB/octave) and the radiation of sound from the mouth (+6 dB/octave). Because the latter is constant across speaking conditions, variation in the combined source+radiation spectrum can be attributed to the source.

Figure 4.19 illustrates the contributions under this model of the different components for a hypothetical vowel with formants at 500, 1500, and 2500 Hz. The individual components of the model are shown in the upper half of the figure, and the spectrum yielded by summing them is shown in the lower half. In this figure, as in the main analysis, default formant bandwidths of 30 Hz for $F1$, 80 Hz for $F2$, and 150 Hz for $F3$.

Figure 4.20 illustrates the measurement on an actual token (the accented vowel of "father", spoken aloud by speaker 5).

The spectral tilt measure is derived from the A1 and A2 measures (peak amplitudes of F1 and F2 respectively) from the smoothed actual spectrum and from the modelled spectrum, as specified in equations 4.8, 4.9, and 4.10:

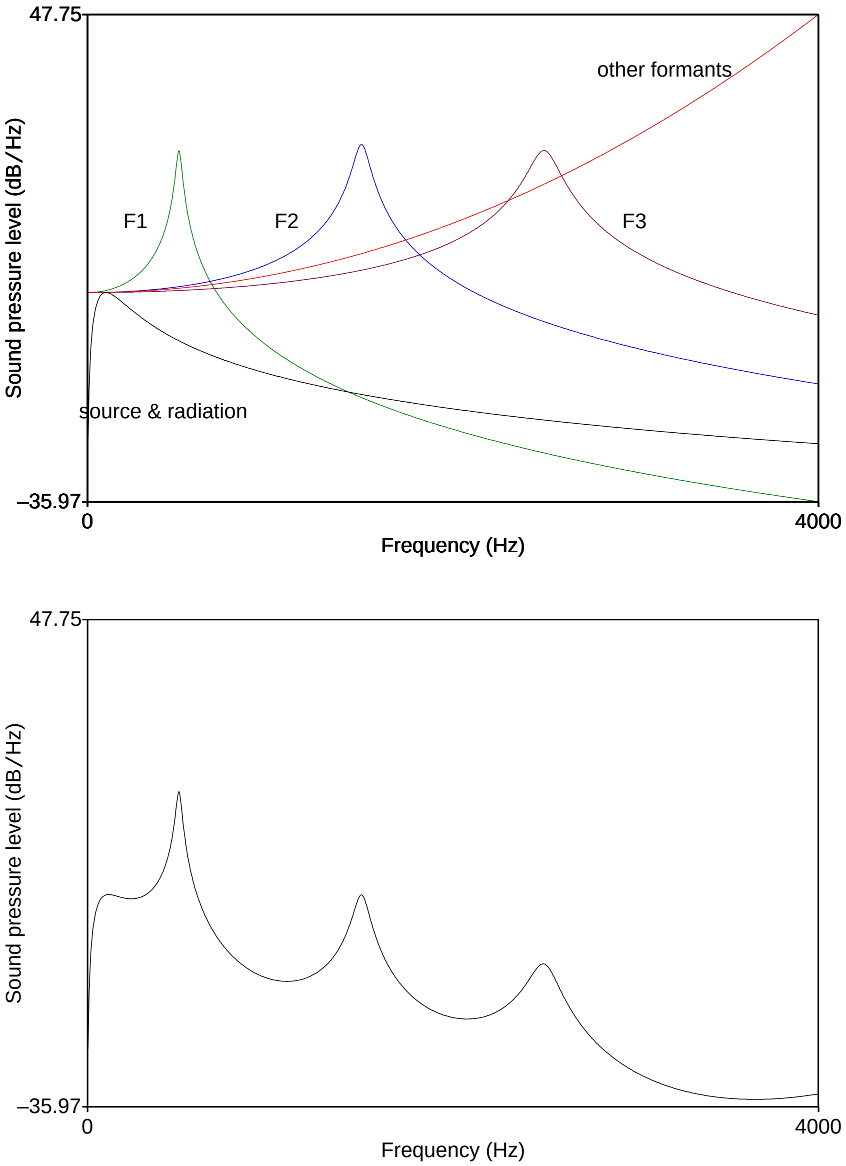


Figure 4.19: Model of spectrum from Fulop et al. (1998), derived from Fant (1960). Upper half shows individual components of model; lower half shows complete spectrum.

$$\begin{aligned}
\text{diff}_{measured} &= A1_{measured} - A2_{measured} && (4.8)\\
\text{diff}_{model} &= A1_{model} - A2_{model} && (4.9)\\
\text{spectral tilt} &= \text{diff}_{measured} - \text{diff}_{model} && (4.10)
\end{aligned}$$

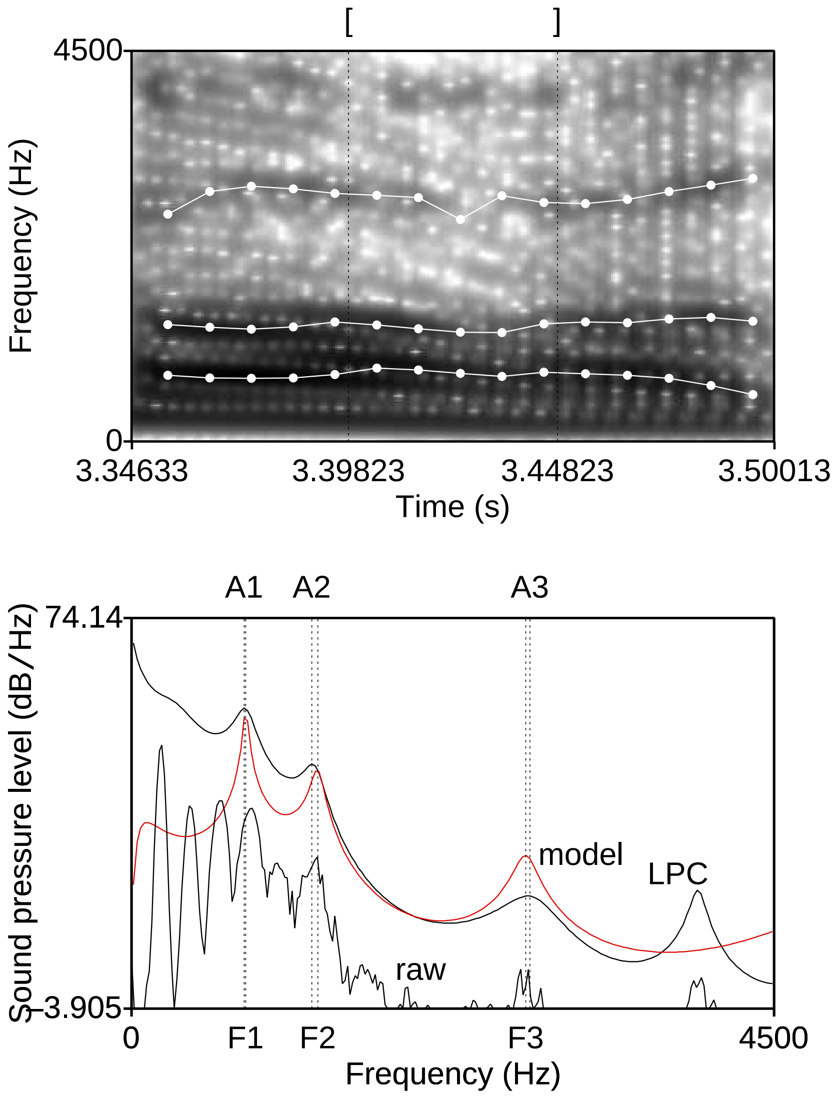


Figure 4.20: Measurement of spectral tilt on the accented vowel of spoken "father" (speaker 5). Upper frame gives spectrogram with 8 ms window, with formant tracks displayed and windowed selection indicated by square brackets at top of figure. Lower frame gives raw spectrum for windowed selection, LPC smoothed spectrum (black), and modelled spectrum based on formant averages from top frame (red).

This measure, by accounting for the effects of formants and other supraglottal contributors to the spectrum, allows us to say something specifically about the source spectrum. Also, because the formants are explicitly modelled, the measure has the potential to be used to compare different vowels.

4.5.5 Summary of existing spectral tilt measures

The current study requires a measure of spectral tilt that can be meaningfully compared across different vowel contexts, that tells us about the source spectrum rather than primarily about the filter properties, and that can be applied to whispered as well as to voiced vowels.

Of all the measures identified in the literature, the only one that fulfils all of these needs is the components model introduced by Fulop et al. (1998). This measure is used in the contrastive emphasis study presented in chapter 5, with a minor modification described in the following section.

4.5.6 Modified optimal measure

The spectral tilt measure presented by Fulop et al. (1998) is the most suitable existing measure in the literature for our purposes, because it is most explicitly designed to exclude all supraglottal influences on the observed spectrum. However, there remains one problem. If F1 and F2 are closer together (as for [ɔ]), the spectral tilt measure will tend to be smaller; if they are further apart (as for [i]), the measure will tend to be larger, even if the glottal slope (the value we seek to represent with our measure) is the same.

In order to control for this, we scale the spectral tilt by the distance between F1 and F2. The slope of the modelled source spectrum, as modelled in equation 4.7, has a relatively constant downward slope of about 6 dB/octave. It seems reasonable to therefore scale the spectral tilt measure by the octave (log-frequency) distance between F1 and F2. Equations 4.11 through 4.13 show how the frequency-scaled spectral tilt measure is derived from the spectral tilt value yielded by equation 4.10 above. The values $F1_{oct}$ and $F2_{oct}$ give the log-scaled F1 and F2 frequencies (in octaves relative to 100 Hz).

$$F1_{oct} = \ln\left(\frac{\frac{F1}{100}}{\ln 2}\right) \tag{4.11}$$

$$F2_{oct} = \ln\left(\frac{\frac{F2}{100}}{\ln 2}\right) \tag{4.12}$$

$$\text{spectral tilt}_{scaled} = \frac{\text{spectral tilt}}{F1_{oct} - F2_{oct}} \tag{4.13}$$

As an example of the range of values yielded by this procedure, consider the spectra illustrated in figures 4.21 and 4.22. The spoken vowel illustrated in figure 4.21 has a spectral tilt of -1.61 dB/octave—meaning that the vowel's spectrum is 1.61 dB/octave flatter than the modelled spectrum. The whispered vowel illustrated in figure 4.22 has a spectral tilt of -14.92 dB/octave—notice how much it diverges from the modelled spectrum above F1, compared with the LPC spectrum in figure 4.21. The difference between these two tokens is expected—the turbulent excitation used in whisper has a much more level spectrum than the periodic excitation of modal voicing (Hanson 1997, p474).

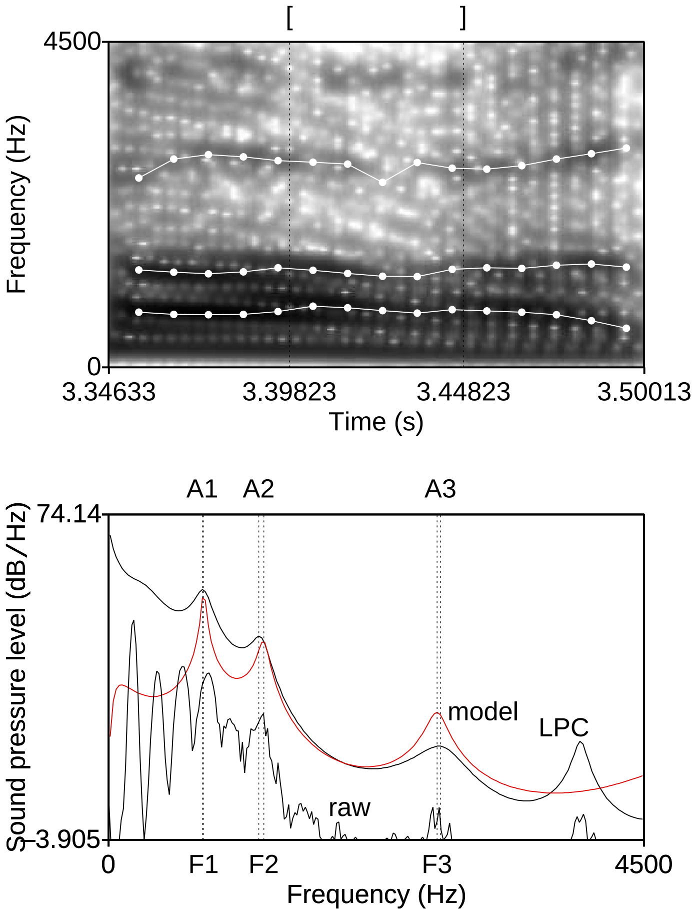


Figure 4.21: Measurement of spectral tilt on the accented vowel of spoken "father" (speaker 5). The upper frame gives the spectrogram with an 8 ms window, with formant tracks displayed and the windowed selection indicated by square brackets at the top of the figure. The lower frame gives the raw spectrum for the windowed selection, the LPC smoothed spectrum, and the modelled spectrum based on formant averages from the top frame. The spectral tilt measure for this token is -1.61 dB/octave.

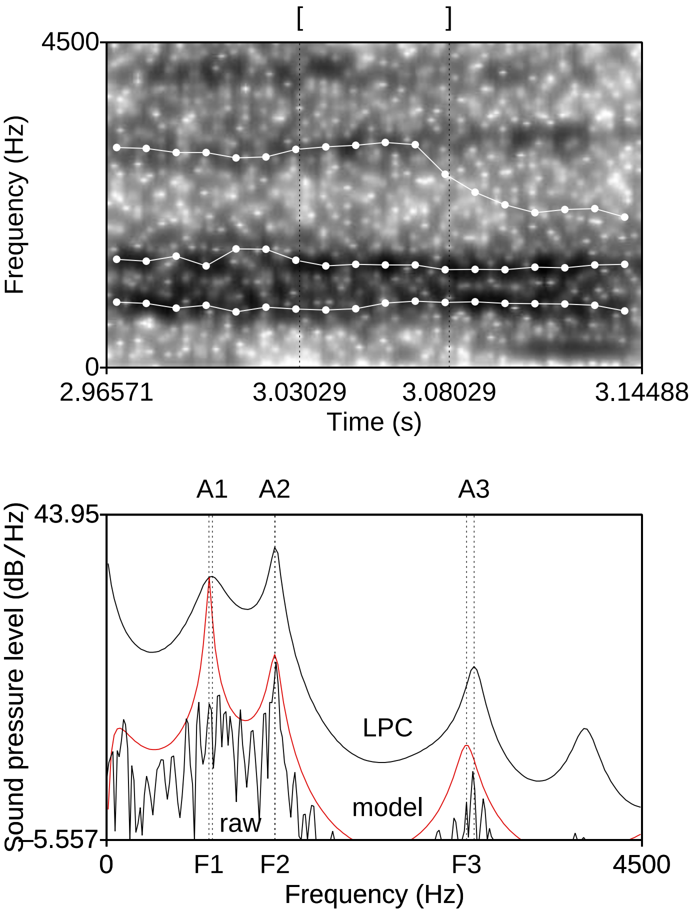


Figure 4.22: Measurement of spectral tilt on the accented vowel of whispered “father” (speaker 5). The upper frame gives the spectrogram with an 8 ms window, with formant tracks displayed and the windowed selection indicated by square brackets at the top of the figure. The lower frame gives the raw spectrum for the windowed selection, the LPC smoothed spectrum, and the modelled spectrum based on formant averages from the top frame. The spectral tilt measure for this token is -14.92 dB/octave.

4.5.7 Validation

Using the data from the contrastive emphasis study described in chapter 5, I performed two tests to verify the validity of the modified spectral tilt measure. First, I compared spoken to whispered statements. The empirical and theoretical literature is clear that a whispered glottal source has a more level spectrum than a voiced source. We should therefore expect whispered tokens to have a lower spectral tilt measure than voiced tokens. We expect a similar pattern in the unmodified measure from Fulop et al. (1998).

Second, I compared the spectral tilt of different vowels in the data. The modifications I made are intended to make the measure vowel-independent; if this goal was successful, we should find no significant differences between the spectral tilt values of the different vowels. This assumes that the glottal source does not systematically covary with vowel quality.

I fitted a mixed-effects linear regression model to the data, with phonation mode (voiced or whispered) as a fixed factor, and with speaker and word as random factors. Separate models were generated for the modified measure and for the original Fulop et al. measure.

Our modified spectral tilt measure yields an average of -7.0 dB/octave for voiced tokens (s.d. = 2.2), and -12.1 dB/octave for whispered tokens (s.d. = 2.2). This difference, illustrated in figure 4.23, is significant ($p < 0.001$).

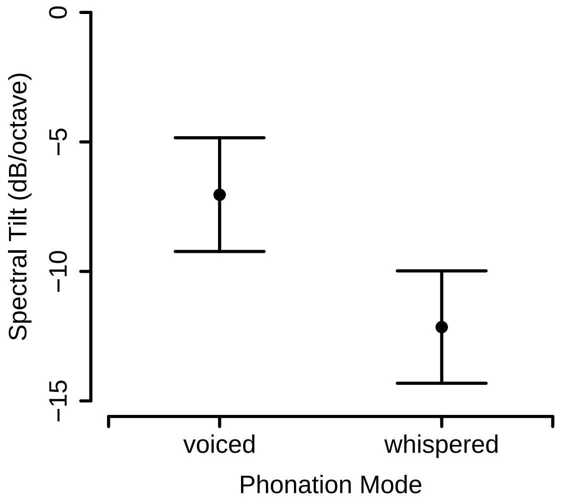


Figure 4.23: Spectral tilt of voiced and whispered vowels using measure modified from Fulop et al. (1998).

A similar pattern is obtained from Fulop et al.'s unmodified measure. Voiced vowels average -5.4 dB (s.d. = 1.2) and whispered vowels average -9.9 dB (s.d. = 1.1). This difference (figure 4.24) is significant ($p < 0.001$).

For the second validation test, I fitted a mixed-effects linear regression model to the data, with word (eight levels) as a fixed factor, and with speaker as a random factor. Separate models were generated for the modified measure and for the original Fulop et al. measure.

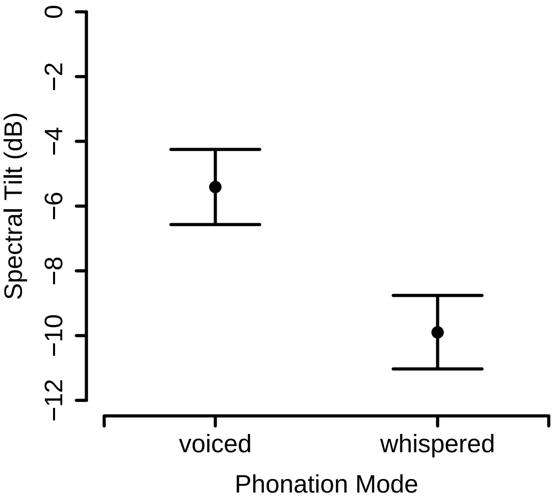


Figure 4.24: Spectral tilt of voiced and whispered vowels using measure from Fulop et al. (1998).

As expected, the original measure from Fulop et al. (1998) showed a great deal of variation between vowels (figure 4.25). Similar vowels, such as the [a] in "fastened", "fastest", and "father", pattern together. Also, some dissimilar vowels, such as the [a] in "marshes" and the [ɔ] in "sources", seem to pattern together. However, overall there is a great deal of variation between vowels. This measure is clearly not vowel-independent.[5]

Fitting the same model to the revised measure, we see a significant change in the relationships between vowels (figure 4.26). Unfortunately, we do not see vowel-independence. There is still a large amount of variation between different vowel qualities.

There are two obvious possibilities for why the normalized spectral tilt measure does not give a vowel-independent value. The first possibility is that the glottal source spectra that produce different vowels do in fact have different slopes. Just as vowels have different intrinsic f_0 (Whalen & Levitt 1995), perhaps they have different intrinsic spectral tilt. However, the magnitude of difference in spectral tilt measures is very large. The difference between the lowest vowel measure ([a]

[5] Note that Fulop et al. use the measure to distinguish between [+ATR] and [-ATR] vowels—they are clearly aware that the measure isn't vowel-independent. Guion et al. (2004) also uses their measure to distinguish vowels from one another. The fact that it is not vowel-independent is not a critique of their use of the measure; simply a caution in the current use.

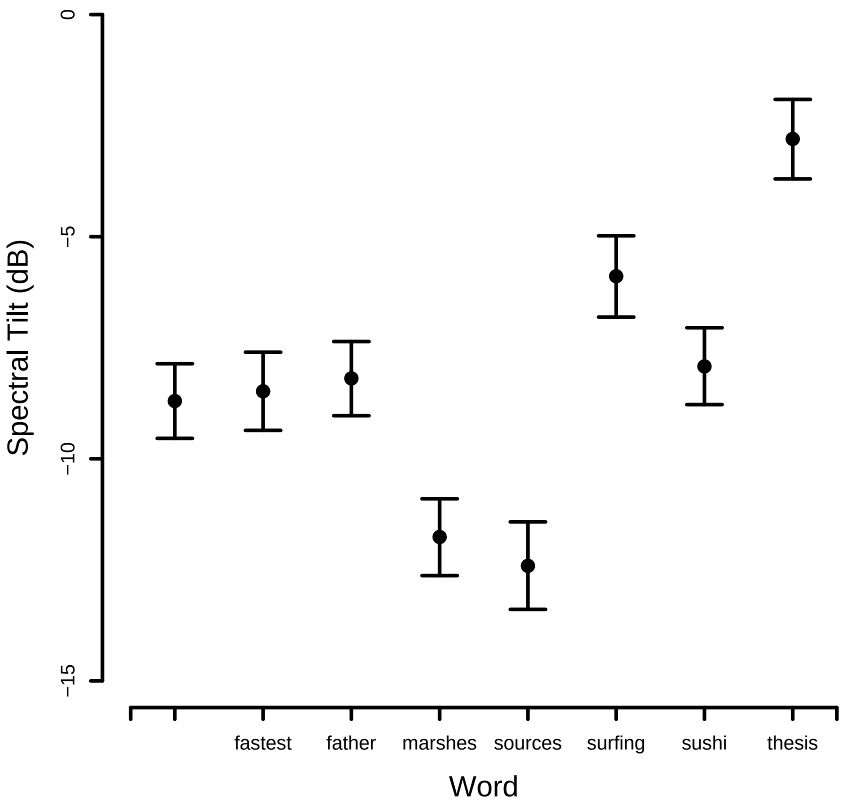


Figure 4.25: Spectral tilt of different vowels using measure from Fulop et al. (1998).

in "marshes") and the highest ([i] in "thesis") is 16.6 dB/octave. For comparison, the difference between voiced and whispered vowels is only 5.1 dB/octave.

The other, more likely possibility is that we have failed to adequately adjust for the effects of the supraglottal resonances in our measure. The original acoustic model from Fant (1960) is already an abstraction, and so will not perfectly model the contributions to the recorded acoustic signal. In addition, we made two abstractions from Fant's model in our calculations. First, we only explicitly modelled the first three formants. In principle, we could measure all formants for a more accurate spectrum.

Second, we did not measure formant bandwidths. We used the default values suggested by Fant (1960) and employed by Fulop et al. (1998). This certainly affects our spectral tilt measure in whisper, where formant bandwidths are known to be greater than they are in voiced speech. It probably also affects

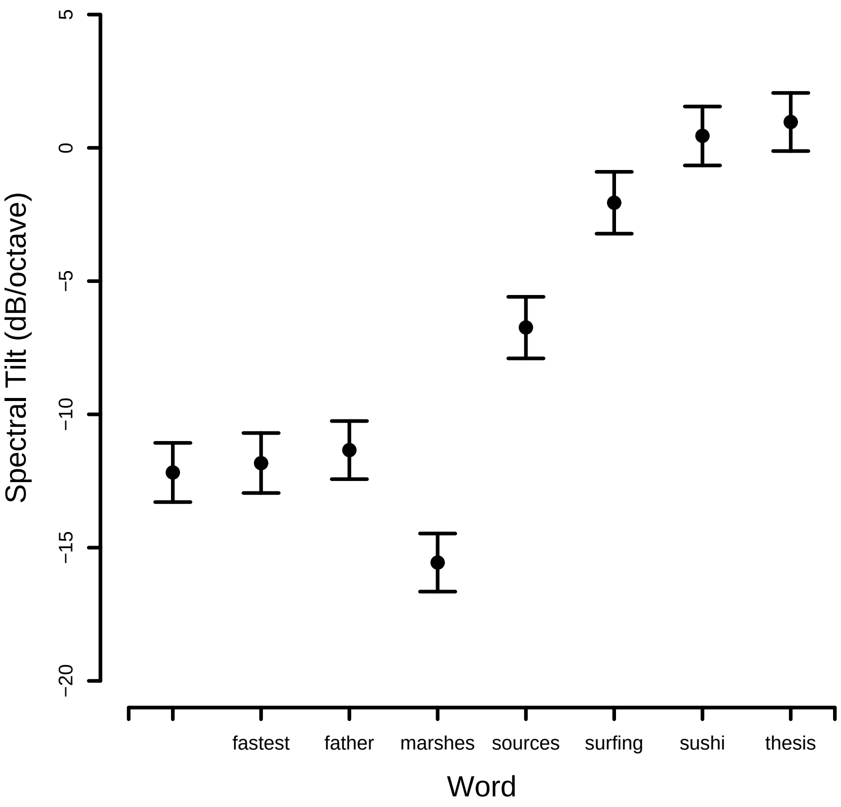


Figure 4.26: Spectral tilt of different vowels using measure modified from Fulop et al. (1998).

inter-speaker comparisons, as speakers vary in the amount of breathiness even in voiced speech (Hanson 1997, Hanson & Chuang 1999, Hanson, Stevens, Kuo, Chen & Slifka 2001).

CHAPTER 5

Contrastive accent study

5.1 Experimental question

The main aim of this dissertation is to examine the control variables that mediate speech motor behaviour. In this experiment, I ask how the parameters that contribute to a prosodic contrast interact, and how their interaction relates to the interactions seen in segmental contrasts.

In example 5.1, the word "surfing" has phrasal stress but is not contrasted with anything. In example 5.2, it has phrasal stress and is contrasted with "swimming" in the preceding sentence.

$$\text{They're away today. They're going surfing.} \tag{5.1}$$

$$\text{They're not going swimming. They're going surfing.} \tag{5.2}$$

When spoken, the accent on "surfing" in 5.2 is made more prominent to express this contrast—it has greater duration and an exaggerated pitch accent (Fry 1955, 1958, Cooper et al. 1985, Eady & Cooper 1986, Pell 2001, Braun 2004, Liu & Xu 2007), and greater amplitude (Traunmüller & Eriksson 2000) or flatter spectral tilt (Sluijter & van Heuven 1996, Remijsen 2001) than the instance of "surfing" in 5.2.

In the current experiment, I use two methods for perturbing the contribution of f_0 to the signal of relative prominence: whispered speech and question intonation.

Whispered speech removes f_0 information completely from the acoustic signal. If there is a single control variable mediating all of the acoustic parameters that signal contrastive context, such a manipulation should cause compensation in one or more of the other acoustic parameters. If, instead, the multiple parameters are controlled by separate control variables, no compensation is expected from the other parameters to the manipulation of f_0.

The results of the endoscopy study in chapter 3 show that normal glottal articulations are not always altered in whisper. If the articulatory contributions to pitch accent, such as vocal fold tension and subglottal pressure (Ní Chasaide & Gobl 1997), are likewise preserved in whispered speech then we might not expect compensation at all, even if the different acoustic parameters are governed by a single control variable. Therefore, a second manipulation was performed with a more directly-observable effect on f_0 production. Following Eady & Cooper (1986), I chose to use question intonation as well as whisper. Examining focus location and intonation, Eady & Cooper (1986) report that in sentence-final accented words, f_0 is a stronger signal of focus (broad versus narrow) in statements than in questions. The difference is slight, but it suggests that question intonation may interfere with the use of f_0 to differentiate focus types. I therefore decided to elicit accents with contrastive and non-contrastive contexts in questions as well as in statements. Because f_0 can be measured acoustically in both statements and questions, this manipulation does not carry the uncertainty that the whisper manipulation carries.

5.2 Speech material

Recordings of sentence pairs like those in 5.1 and 5.2 above give us information about how duration, f_0, intensity, and spectral tilt are used to signal contrastive context. Because the sentence in which the target word is found does not change between these two conditions, we can assume that any differences in these measures are due solely to the difference in context.

See table 5.1 for all target sentences; appendix C gives the complete list of sentences used in all manipulations, including the context sentences used to elicit contrastive and non-contrastive readings. Speakers' productions of these sentences give us a baseline for how each acoustic parameter is used to signal contrastive context. Next, we introduce our experimental manipulation.

Their seatbelts will be fastened.
She can cycle fastest.
They're golfing with their father.
They're searching the marshes.
It's on in the morning.
He checked his main sources.
They're going surfing.
They ordered some sushi.
She's writing a thesis.
He's very thirsty.

Table 5.1: Sentences used to establish baseline acoustic contributions to signalling contrastive context.

The sentences listed in table 5.1 were elicited in whisper as well as in normal voice. Whispered speech, lacking vocal fold vibration, lacks f_0 as an acoustic parameter to signal contrastive context. If the other acoustic parameters are governed by the same control variable, we would expect them to compensate in whispered speech by showing greater differences between non-contrastive and contrastive contexts.

This manipulation is somewhat reliant on the assumption—currently untested—that the articulatory gestures which generate f_0 contours in normal speech are not present in whisper. The results from the endoscopy investigation of obstruents (chapter 3) suggest that this assumption may be false. If we do not see compensation in the current study, then, it may be due either to the operation of separate control variables for the different acoustic parameters, or to the fact that the manipulation did not in fact alter the articulatory behaviour regarding f_0. If we *do* see compensation under the whisper manipulation, it will strongly suggest both that the manipulation was successful in inhibiting the articulatory contribution to f_0 and that there is a single control variable responsible for f_0 and whichever acoustic parameters exhibit compensation.

Because of the uncertainty about whether the whisper manipulation actually removes articulations underlying f_0 manipulations, a second manipulation was also performed. The f_0 environment at the end of a sentence is different for a yes/no question than for a statement: typically, questions end with a high boundary tone, while statements end with a low boundary tone. I elicited target words sentence-finally in questions which were structurally almost identical to the statements given above—compare 5.1 and 5.2 above with 5.3 and 5.4:

They're away today. Are they going surfing? (5.3)

They're not going swimming. Are they going surfing? (5.4)

Because of the high boundary tone in questions, speakers have a more limited pitch range available with which to signal contrastive context (Grabe, Post, Nolan & Farrar 2000). This claim is tested for the current data in section 5.6.12 below. With the f_0 parameter thus attenuated, we can look for increased magnitude of one or more other parameters signalling contrastive context.

This manipulation has the advantage over the whisper manipulation of allowing us to observe acoustically whether the articulatory control of f_0 has, in fact, been changed. Its disadvantage is that the effect on f_0 is less complete: there is still *some* f_0 difference between tokens in non-contrastive and contrastive contexts; in whisper, if the manipulation is successful, there is no remaining f_0 difference.

5.3 Participants

Recruitment of participants was guided mainly by one criterion: whether they spoke a variety of English in which pitch accents are consistently expressed with high f_0 excursions, rather than a mixture of high and low. This allows us to directly compare the magnitude of pitch accents across conditions. In exploratory recordings, Southern British English speakers were found to fit this requirement. Speakers of other varieties—particularly from the north of England and from Scotland—exhibited high pitch accents in statements, but frequently had low pitch accents in questions. This variation is also observed by Grabe et al. (2000)—they report that speakers from Newcastle show high pitch accent in statements and low pitch accent in questions. They also report low pitch accent in both conditions for Belfast speakers. Speakers from Leeds and Cambridge in their study showed high pitch accents in both conditions.

Four male and two female speakers were recorded for this study, all native speakers of Southern British English.

Speaker	Order
1	S Q W
2	W S Q
3	Q W S
4	S W Q
5	Q S W
6	W Q S

Table 5.2: Speakers and block orders. S = spoken statements; Q = spoken questions; W = whispered statements.

5.4 Recordings

Each context + target sentence pair was made into a slide for presentation on a computer screen.

A first round of pilot recordings demonstrated that, when slides from all prosodic conditions were randomized together in a single block—contrastive and non-contrastive contexts for statement, question, and whispered tokens—speakers found it difficult to use the correct type of accent. They confused the different conditions, producing contrastive accents in non-contrastive contexts and not producing contrastive accents in contrastive contexts. It is likely that there were too many dimensions changing between tokens: the intonational manipulation (statement/question/whisper) and the context variation (contrastive/non-contrastive).

Therefore, the slides were divided into three blocks. One block contained all of the spoken-aloud statement tokens (contrastive and non-contrastive); another contained all of the question tokens (contrastive and non-contrastive), and the third contained all of the whispered tokens (contrastive and non-contrastive).

Within each block, the twenty slides (ten target sentences, each in both a contrastive and a non-contrastive context) were presented five times. Each of the five sets of twenty slides was combined in a unique random order.

In order to prevent order of presentation from confounding the results, the blocks were presented in a different order to each speaker according to a Latin Square design. Each of the six possible orders of spoken statements, spoken questions, and whispered statements was assigned to one of the six speakers. The order assigned to each speaker is presented in table 5.2.

In a second set of pilot recordings, in which tokens were separated into blocks as above, speakers did not consistently produce contrastive accent on the target words when it was indicated by the context sentence. While I did not want to bias speaker behaviour by overly-specific instruction, it was crucial for this experiment that a consistent difference between contrastive and non-contrastive tokens was produced, at least in the base condition (spoken statements).

In order to remind speakers to produce contrastive accent, without biasing them about *how* to produce it, slides with a non-contrastive context sentence had "(sequence)" prepended to them, to indicate a sequence of two related sentences, and slides with a contrastive context sentence had "(contrast)" prepended to them. Instructions were also given at the start of the recordings to pay particular attention to this difference. Reminders were given periodically throughout the recordings. These additions succeeded in eliciting consistent differences between the contrastive and non-contrastive tokens. Sample slides are presented in figure 5.1.

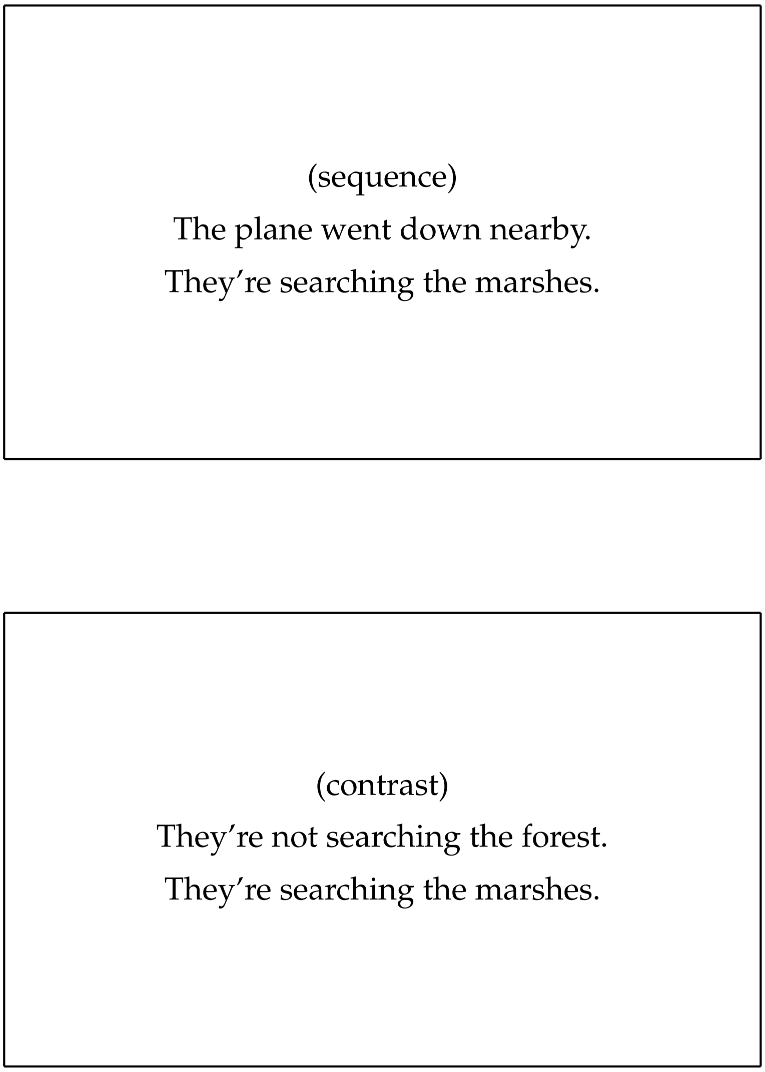
Recordings were made in a sound-treated booth. The experimenter sat with speakers during the recordings. The experimenter controlled the presentation of slides (approximately one half second between the end of reading one slide and presentation of the next). If the experimenter (the author) perceived that the accent was wrong on a given production (contrastive accent when a slide gave a non-contrastive context, or vice versa), the speaker was asked to repeat that token. Instructions were repeated between sets and between blocks if a speaker seemed to be ignoring the contrastive/non-contrastive difference.

Speakers were given water and encouraged to drink whenever needed. In particular, they were advised to drink if they began to sound dry—aside from the importance of attending to participants' well-being, excessive mucosal viscosity (stickiness) due to dryness produces acoustic events which interfere with clear measurements.

Participants were given £5.00 and some chocolate in thanks for their participation.

5.5 Measurements

Five acoustic properties were measured for each token in this dataset: f_0, duration, amplitude, vowel quality, and spectral tilt. See chapter 4 for details on how



(sequence)
The plane went down nearby.
They're searching the marshes.

(contrast)
They're not searching the forest.
They're searching the marshes.

Figure 5.1: Sample slides presented to participants.

each acoustic parameter was measured.

5.6 Results

In this experiment, I ask whether a reduction in the contribution of one acoustic parameter to signalling contrastive context elicits a compensatory increase in the contributions of any of the other acoustic parameters that signal contrastive context. Before answering the question, we first need to establish three properties of the data:

1. The parameter being manipulated—f_0—is consistently used by speakers to signal whether the context is contrastive or not.
2. At least one other parameter—in this case, one of duration, amplitude, spectral tilt, and vowel quality—is also used to signal the type of context.
3. The manipulations used (statements versus questions, and spoken versus whispered) successfully alter the extent to which f_0 contributes to signalling the type of context.

In section 5.6.2, I examine the spoken statement data. I find that f_0, duration, and amplitude all show a significant effect of contrastive context in the expected direction. Spectral tilt shows no effect. This satisfies prerequisites 1 and 2 above. Section 5.6.8 presents the whisper manipulation. In this manipulation, it is assumed that an f_0 manipulation was successful. Duration and amplitude show no evidence of compensation. Section 5.6.11 presents the question manipulation, showing that f_0 was successfully reduced as a signal of contrastive context. Neither duration nor amplitude exhibit compensation. Instead, they show the reverse: when the contribution of f_0 is reduced, so are the contributions of duration and amplitude.

Section 5.7 presents some discussion of the results and how they relate to the purpose of the study.

5.6.1 *Statistics used*

As with the endoscopy study, I used mixed-effects linear models to test for significant effects in the data (see section 3.5.11). I use the *lmer* (Linear Mixed Effect Regression) function from the `lme4` package Bates (2005) in the R statistical software (R 2008). I report p-values derived from Markov Chain Monte Carlo (MCMC) sampling (10 000 samples), which avoids the anticonservative bias Baayen (2008) reports for the traditional *t*-statistic when dealing with small sample sets. (The current data is a small data set, with only 554 measured data points across all speakers and conditions.)

A recent analysis (Quené & van den Bergh 2008) shows that repeated measures techniques where there are multiple crossed random factors are at greater risk of Type I error (underestimating the probability of the null hypothesis) than the alternative mixed-effects models. This is relevant for the current study, where there are two random factors (speaker and word). For further discussion of the theory behind mixed-effects models, see Neter et al. (1996), Bates (2005), and Baayen (2008, chapter 7).

Because I am using MCMC sampling to determine p-values, there are no F-statistics to report. I include graphs showing group variances from the data to give the reader a visual idea of the effects reported. Using the raw variance across the full dataset would tend to obscure true effects, as between-speaker and between-word differences would be added to true within-condition variances. I therefore applied the following adjustment to data for each variable before plotting the graphs. For each speaker, I calculated the difference between that speaker's mean and the grand mean of the data. Each data point from that speaker was shifted by that difference, so that the speaker's mean equalled the grand mean. This adjustment preserved differences between conditions (ie, between contrastive and non-contrastive tokens), but removed inter-speaker differences. This is akin to how the statistical test deals with inter-speaker variation. The same adjustment was then applied to the other random factor (target word). This manipulation mirrors the adjustments for between-group effects of random variable performed in the statistical analyses.

5.6.2 Basic effects

The following sections present the tests for a main effect of contrastive context on the final accented syllable in spoken statements. The data show that f_0, duration, and amplitude all signal contrastive context in spoken statements, but vowel quality and spectral tilt do not. All tests for main effects use a mixed-effects ANOVA with contrastive context as a fixed factor (2 levels: contrast or no contrast) and random factors word (9 levels) and speaker (6 levels).

5.6.3 Main effect on f_0

In spoken statements, words in a non-contrastive context had a peak 0.72 semitones lower than the preceding pitch accent, on average. Words in a contrastive context averaged 0.37 semitones higher than the preceding pitch accent. The difference between accents in contrastive and non-contrastive contexts relative to the prenuclear accent was 1.03 semitones—a significant difference ($p < 0.001$). Figure 5.2 illustrates the difference.

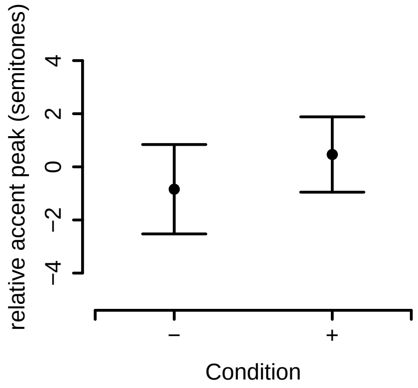


Figure 5.2: Mean f_0 nuclear accent peak relative to prenuclear accent in non-contrastive (-) and contrastive (+) contexts, using a logarithmic (semitone) scale. Bars indicate one standard deviation above and below the mean.

5.6.4 *Main effect on duration*

In order to determine whether any duration effect was confounded by varying speech rate, a test was made with the control duration as a dependent variable. This test showed a significant effect ($p < 0.001$), but as figure 5.3 shows, the effect was the reverse of that seen above for the accented syllable: the control duration (which received pre-nuclear accent) was 16 ms (6%) shorter with contrastive context than with non-contrastive context. The effect of contrastive context on the target word duration is thus very unlikely to be due to varying speech rate. We can therefore conclude that any positive effect of contrastive context on the accented syllable duration is genuine.

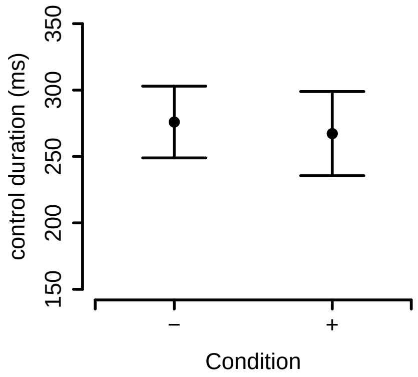


Figure 5.3: Mean duration of control word in non-contrastive (-) and contrastive (+) contexts. Bars indicate one standard deviation above and below the mean.

A significant effect of contrastive context on accented syllable duration was seen ($p < 0.001$): syllables in a contrastive context averaged 16.8 ms (8%) longer than those in a non-contrastive context. This effect is shown in figure 5.4. It is not a large effect: studies on perception find 50% discriminability thresholds between 5% and 10%, depending on the task (Fujisaki, Nakamura & Imoto 1975, Quené 2007). The current result—syllable duration increasing 6% in contrastive context over non-contrastive—lies at the border of perceptibility. Numerically, it is statistically significant, suggesting that this is a consistent part of the motor control involved in signalling contrast.

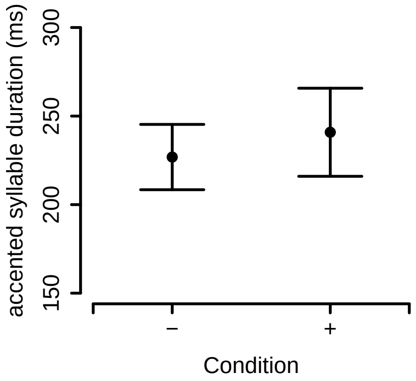


Figure 5.4: Mean duration of stressed syllable in non-contrastive (-) and contrastive (+) contexts. Bars indicate one standard deviation above and below the mean.

The effect carried over onto the unstressed final syllable of the word as well—it averaged 18 ms (6%) longer in contrastive than non-contrastive contexts ($p < 0.001$). Figure 5.5 illustrates this effect.

5.6.5 *Main effect on amplitude*

There was a significant effect of contrastive context on the peak amplitude of the accented vowel ($p < 0.001$): accents in a contrastive context had an average peak amplitude 1.14 dB higher than accents in a non-contrastive context (figure 5.6).

The just noticeable difference for isolated vowel sounds is around 1.2 dB (Flanagan 1955). The perceptual threshold for differences in a natural speech context is likely to be even higher, as there are many other acoustic fluctuations in speech that listeners must also attend to. Our observed difference of 1.14 dB is therefore

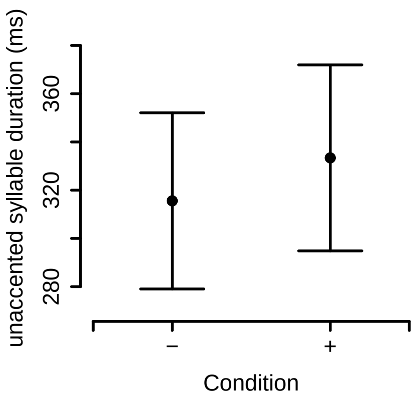


Figure 5.5: Mean duration of unstressed syllable in non-contrastive (-) and contrastive (+) contexts. Bars indicate one standard deviation above and below the mean.

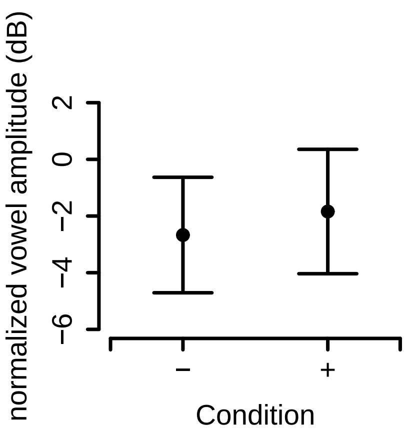


Figure 5.6: Mean amplitude of stressed vowel in non-contrastive (-) and contrastive (+) contexts. Bars indicate one standard deviation above and below the mean.

unlikely to be perceptually useful. As with duration, though, the tendency is statistically significant, suggesting a real difference in articulation.

5.6.6 *No effect on vowel quality*

Two tests were performed for each of F1 and F2: one with vowels having a lower average frequency of that formant, and one with vowels having a higher average frequency of that formant. They are illustrated in table 5.3. None of the tests showed a significant effect of context: vowels in a contrastive context were not less centralized than vowels in a non-contrastive context.

	Vowels	p	
Set:		linear	log
F1 low	/i ɔ u/	0.341	0.384
F1 high	/æɑ/	0.850	0.580
F2 low	/æɑ ɔ/	0.080	0.077
F2 high	/i u/	0.205	0.232

Table 5.3: Effects of contrastive context on formant frequencies.

5.6.7 *No effect on spectral tilt*

As seen in figure 5.7, there was no significant effect of contrastive context on spectral tilt (p=0.233).

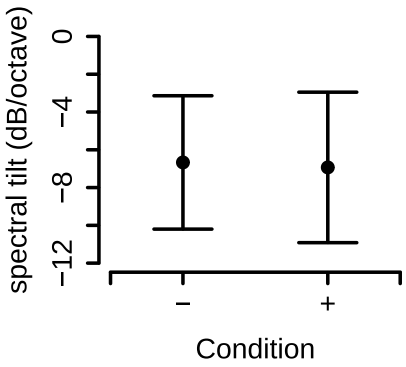


Figure 5.7: Mean spectral tilt measures in non-contrastive (-) and contrastive (+) contexts. Bars indicate one standard deviation above and below the mean.

5.6.8 *Responses to the whisper manipulation*

In the first manipulation, I elicited the same statements in whisper as in normal speech. With whisper, the presence of f_0 as an acoustic cue to contrastive context is completely eliminated. Whether this manipulation successfully eliminated the articulatory adjustments corresponding to f_0 manipulation cannot be directly answered with the present data (see the discussion at the end of section 5.2 above).

The tests reported in the following sections are based on a mixed-effects linear model with context (non-contrastive or contrastive) and phonation mode (spoken or whispered) as fixed factors, and with word (9 levels) and speaker (6 levels) as random factors. The dependent (response) variables tested are duration (stressed syllable, unstressed syllable, control duration) and amplitude.

5.6.9 Duration

An examination of the control duration (figure 5.8) shows no confounding influence of speech rate on on the following effects. There is an effect of context ($p = 0.006$): the control duration averages 14 ms (5%) *shorter* in contrastive than in non-contrastive contexts. There is a significant effect of phonation mode ($p < 0.001$): control durations are 15 ms (5%) greater in whispered than in normal speech. There is no interaction between context and phonation mode (p=0.680).

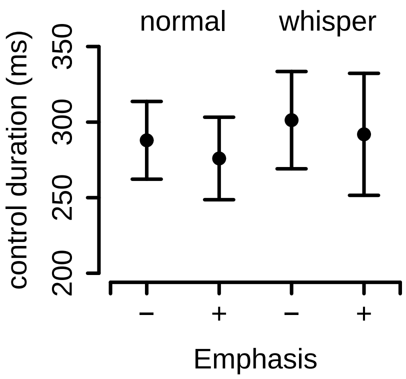


Figure 5.8: Control duration in non-contrastive (-) and contrastive (+) contexts in normal and whispered speech. Bars indicate one standard deviation above and below the mean.

On the accented syllable of the target word, there is a main effect of context ($p < 0.001$): syllables with a contrastive context average 17 ms (8%) longer than those without in spoken statements, and 13 ms (5%) longer in whispered statements (figure 5.9). There is a main effect of phonation mode: syllables are significantly longer when whispered (by an average of 11 ms, 5%) than when spoken normally ($p < 0.001$). There is no interaction between the effect of context and that of phonation mode (p=0.251)—no compensation.

Similar patterns are seen on the unaccented syllable (figure 5.10). There is a main effect of context ($p < 0.001$)—17 ms (5%) in normal speech and 10 ms (3%) in whispered speech. Unlike on the previous syllable, there is no main effect of phonation mode ($p = 0.773$). There is no interaction between context and phonation mode ($p = 0.264$).

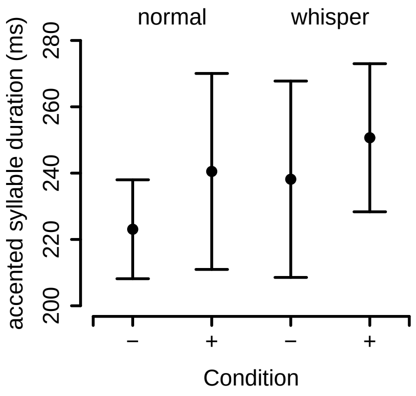


Figure 5.9: Mean duration of accented syllable in non-contrastive (-) and contrastive (+) contexts in normal and whispered speech. Bars indicate one standard deviation above and below the mean.

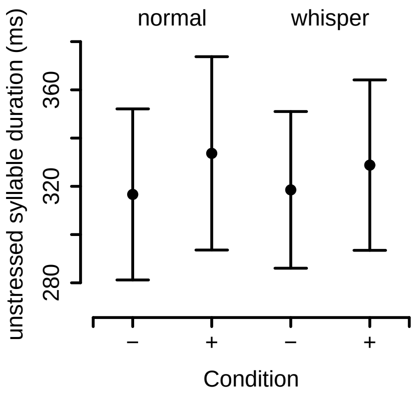


Figure 5.10: Mean duration of unaccented syllable in non-contrastive (-) and contrastive (+) contexts in normal and whispered speech. Bars indicate one standard deviation above and below the mean.

5.6.10 *Amplitude*

We find an overall main effect of context on peak amplitude ($p = 0.046$): peak amplitude is 1.21 dB greater in a contrastive than in a non-contrastive context. There is also a main effect of phonation mode ($p < 0.001$): peak amplitude is 2.36 dB greater in whisper than in normal speech. (Remember that the peak amplitude is measured relative to the peak amplitude in the accented syllable of the control sequence, so our measure will not reflect the fact that whispered speech overall has a lower amplitude than modal speech.) There is no interaction between the effect of context and the effect of phonation mode ($p = 0.347$).

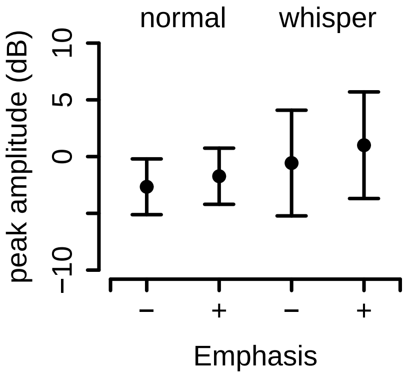


Figure 5.11: Mean amplitude in non-contrastive (-) and contrastive (+) contexts in normal and whispered speech. Bars indicate one standard deviation above and below the mean.

5.6.11 *Responses to the question manipulation*

The following sections report results from the question manipulation, in which the contribution of f_0 to signalling contrastive context was reduced by placing the words in yes/no questions, near the high boundary tone. Neither of the other acoustic parameters that signal contrastive context show compensation. Interestingly, the duration measure mirrors the f_0 measure in showing a weaker effect of contrastive context in questions than in statements.

All tests reported in this section are based on a mixed-effects linear model with context (non-contrastive or contrastive) and sentence type (statement or question) as fixed factors, and with word (9 levels) and speaker (6 levels) as random factors.

5.6.12 *Fundamental frequency*

Using f_0 peak as a dependent variable, the main effect of context seen in the statement data remains ($p < 0.001$): overall, tokens in a contrastive context have an f_0 peak 0.52 semitones higher than tokens in a non-contrastive context, relative to the prenuclear accent. Crucially for this experiment, we also find a significant interaction between context and sentence type ($p < 0.001$). In statements, accent peaks averaged 1.33 semitones higher in contrastive than in non-contrastive contexts, relative to the previous accent. In questions, the difference between accent peaks in contrastive contexts and those in non-contrastive contexts was only 0.49 semitones.

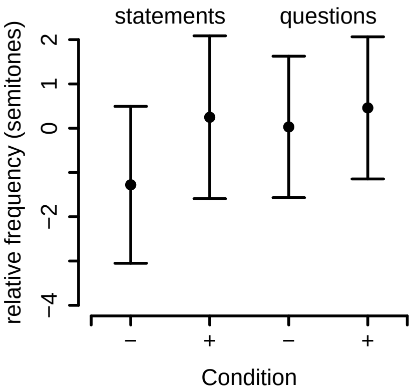


Figure 5.12: Mean f_0 peak relative to prenuclear accent in non-contrastive (-) and contrastive (+) contexts in statements (st) and questions (qu). Bars indicate one standard deviation above and below the mean. Frequency is shown with logarithmic scaling (semitones).

Our manipulation of the f_0 contribution to signalling contrastive context was successful: f_0 contributes to more in statements than it does in questions (figure 5.12).

Also, the difference between statements and questions is significant. The pitch accent on the final word of questions averaged 0.77 semitones higher, relative to the preceding accent, than the pitch accent on the final word of statements ($p < 0.001$). As mentioned in section 4.2, this does not mean that questions had a rising intonation at the end; it simply means that the final pitch accent was higher in questions than in statements (with all other pertinent intonational variables kept constant).

5.6.13 Duration

When examining spoken statements in section 5.6.4, we saw that the duration of the accented syllable showed an effect of context.

Looking at statements alongside questions now, we see a main effect of contrastive context ($p < 0.001$): tokens in a contrastive context have an 11 ms longer accented syllable than those in a non-contrastive context. We also see a significant interaction between contrastive context and sentence type ($p < 0.001$). Looking at the marginal means, we see that contrastive context increases accented syllable duration by an average of 19 ms (9%) in statements, but only by 3 ms (1%) in questions, over its duration in a non-contrastive context. This is shown in figure 5.13.

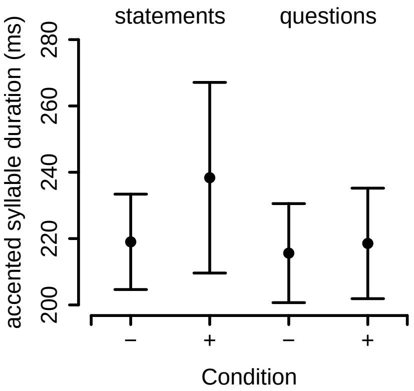


Figure 5.13: Mean accented syllable duration in non-contrastive (-) and contrastive (+) contexts in statements and questions. Bars indicate one standard deviation above and below the mean.

The same pattern was observed in the final (unaccented) syllable: an overall effect of contrastive context ($p < 0.001$), interacting with sentence type ($p = 0.020$): there is a 20 ms (6%) effect of contrastive context in statements, but only a 3 ms (1%) effect in questions.

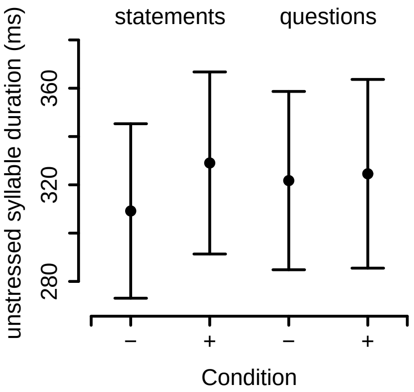


Figure 5.14: Mean unaccented syllable duration in non-contrastive (-) and contrastive (+) contexts in statements and questions. Bars indicate one standard deviation above and below the mean.

Again, we check these results against the pattern of the control duration across the conditions (figure 5.15). We find a significant main effect of contrastive context on the control duration ($p < 0.001$)—as in section 5.6.4. The control duration averages 8 ms (3%) shorter in tokens with a contrastive context than in tokens with a non-contrastive context. There is a main effect of sentence type—control durations are 13 ms (5%) shorter on average in questions than they are in statements ($p < 0.001$). There is an almost-significant interaction of context and sentence type ($p = 0.057$)—the negative effect of contrastive context tends to be greater in statements than in questions. These patterns demonstrate that the effects observed above in the target words are not an artefact of varying speech rate.

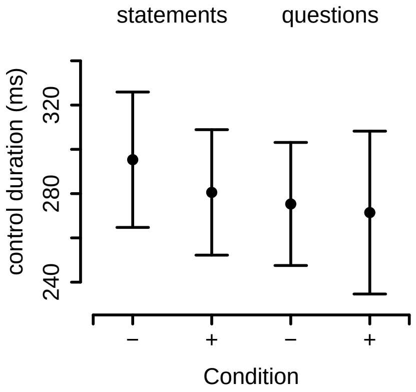


Figure 5.15: Mean control duration in non-contrastive (-) and contrastive (+) contexts in statements and questions. Bars indicate one standard deviation above and below the mean.

5.6.14 Amplitude

In section 5.6.5, we saw that the base condition (spoken statements) shows a small but statistically significant effect of contrastive context on peak amplitude.

Looking at statements and questions together, we find a main effect of contrastive context ($p < 0.001$). In statements, vowels in a contrastive context are 1.11 dB higher than those in a non-contrastive context; in questions, vowels in a contrastive context are only 0.13 dB higher than those in a non-contrastive context. There is a significant effect of sentence type ($p = 0.023$); questions have a 0.07 dB higher amplitude than statements. There is a significant interaction between context and sentence type ($p = 0.020$)—the effect of contrastive context on amplitude is significantly greater in statements than in questions (figure 5.16).

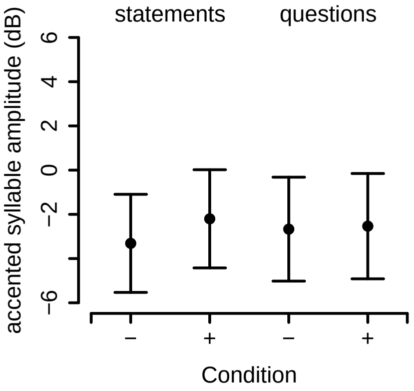


Figure 5.16: Mean amplitude in non-contrastive (-) and contrastive (+) contexts in statements and questions. Bars indicate one standard deviation above and below the mean.

In other words, like its duration and its f_0 peak, the amplitude of the accented syllable is a statistically stronger signal of contrastive context in statements than in questions. The difference remains below the JND of 1.2 dB for vowel sounds observed by Flanagan (1955).

5.7 Summary

The study presented in this chapter provides a phonetic profile of the effects of contrastive context as conveyed in Southern British English, and yields some evidence on the question of single or multiple control variables for prosodic contrasts.

The current data confirm that contrastive context is, indeed, signalled by multiple acoustic cues. We find that the f_0 peak of the pitch accent is higher for accented words in a contrastive context than in a non-contrastive context; the duration of the accented syllable is greater; and the peak amplitude is higher. Vowel quality and spectral tilt do not vary systematically with contrastive context in the current data.

In the whisper manipulation, there was no compensatory behaviour in duration or amplitude. This is consistent with separate control variables for these different measures in the signalling of contrastive context. Remembering that glottal "voicing/devoicing" gestures were observed in the endoscopy study (chapter 3), we cannot discount the possibility that the same is happening in the current study. Perhaps the manipulations of vocal fold tension and subglottal pressure that produce f_0 changes in normal speech are also performed in whispered speech, even though they have no perceptible effect on the acoustic signal. If so, and if changes in articulation are required to trigger compensation, then a control variable governing f_0, duration, and amplitude together would still not be expected to trigger compensation in the whisper condition. Short of electromyographic or endoscopic investigation, we cannot tease apart these two possible interpretations of the results from the whisper manipulation.

In the question manipulation, we were able to manipulate the degree to which f_0 signalled contrastive context. The f_0 difference between accents in contrastive and non-contrastive contexts is much weaker at the end of a question than at the end of a statement. Neither of the other cues identified as signalling contrast—duration and amplitude—exhibited compensation under this manipulation. In fact, both duration and amplitude showed evidence of being attenuated in the same direction—weaker signalling of contrast in questions than in statements. This fits neither the null hypothesis (separate control variables, no response of one parameter to manipulation of another) nor the compensation hypothesis (increased use of one parameter in response to decrease of another). The relationship observed is consistent with the idea that duration and amplitude are not independent of f_0 in production—that speakers cannot physically alter one without affecting the others. However, the work of Berkovits (1984) comparing English and Hebrew sentence-final intonation demonstrates that it is physiologically possible to manipulate f_0 and duration independently of one another. Moreover, what we know of the physiological and acoustic aspects of production suggests that these parameters can be manipulated independently.

So while the whisper manipulation give results consistent with the hypothesis of separate control variables for f_0, duration, and amplitude, the question manipulation gives results that are consistent with neither of the original hypotheses.

Both manipulations give results that differ from the type of results seen in bite block studies of lip closure and vowel height, and in studies of acoustic motor equivalence. As discussed above, the whisper manipulation may have failed to alter f_0 behaviour in such a way as to trigger compensation. However, the question manipulation was clearly successful in setting up the preconditions for compensation. It remains to be seen what the unexpected result says about motor control variables in the signalling of contrastive context. This is discussed further in chapter 6.

CHAPTER 6

Discussion and conclusions

In this dissertation, I inquire about the nature of speech motor control—specifically, motor control variables, around which online adaptation to different speech conditions is organized.

6.1 Voicing contrasts in whispered speech

Chapter 3 presents two experiments probing the utility of whispered speech as a tool for investigating the question of articulatory versus acoustic targets in the production of phonological voicing contrasts.

6.1.1 Articulatory study

In the examination of endoscopic data, I present a procedure for extracting quantitative measures of glottal aperture from the high-dimensional video data. From these measures, we see that there are systematic glottal aperture differences: voiced and voiceless obstruents are produced with the same glottal apertures in whispered speech as in normal speech—this element of the articulatory contrast is preserved intact in whisper.

Readers are reminded that, in this study, we use maximum distance between the vocal folds as our measure of glottal aperture. This can be expected to correlate with other measures of glottal aperture, such as glottal area, but it is not an identical measure. For example, a glottis with the vocal folds parted slightly along their whole length may give the same measure as one with mostly closed vocal folds and a posterior glottal chink (as in most speakers' whispered glottal state), if we were to use the current technique. See figure 6.1, reproduced from figure 2.5 for an illustration of such differences. However, if we were to measure glottal area (perhaps, the number of pixels in the visible glottis), the former would have a greater glottal aperture than the latter. My impression is that, because of the "glottal chink" configuration that dominates in whisper, a study that looked at glottal area rather than glottal opening would show less difference in aperture between normal and whispered speech than the current study does. However, I suspect that the conclusions regarding obstruent contrasts would be substantially similar to those drawn here.

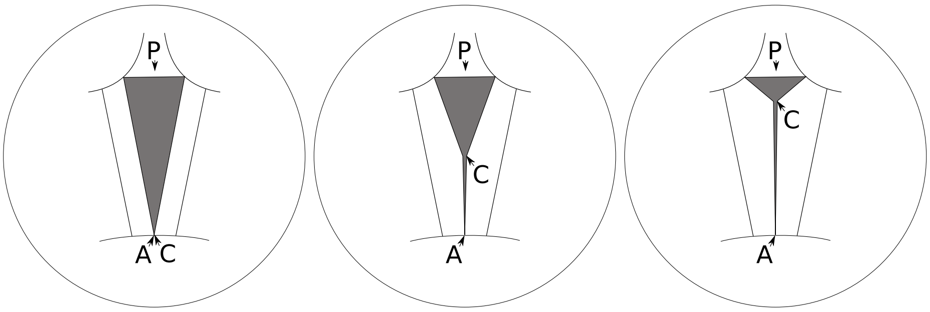


Figure 6.1: Reproduction of figure 2.5, to illustrate the difference between measuring glottal aperture as distance and as area.

The first benefit of this finding is that it provides us with a quantitative characterization of glottal behaviour in whispered voicing contrasts. Figure 6.2 presents the results for spoken and whispered tokens (replicating the left portion of figures 3.28 and 3.35).

We are now able to empirically adjudicate speculations in the literature that heretofore had not been directly tested. Perkins et al. (1976) found that stuttering rates are reduced in whisper relative to normal speech. They attributed this to the reduced coordinative complexity of speech, assuming that no glottal gestures are made in whispered speech. The current results clearly falsify this assumption. On the other hand, the current results largely confirm a long line of assertions in the literature, from Sweet (1877, 1906), through Pike (1943), Malmberg (1963), Abercrombie (1967), and Catford (1964, 1977), to Laver (1994), that in whispered speech, phonologically voiceless consonants should be produced with the voiceless glottal state (as they are in normal speech), and phonologically voiced consonants should be produced with the whisper glottal state

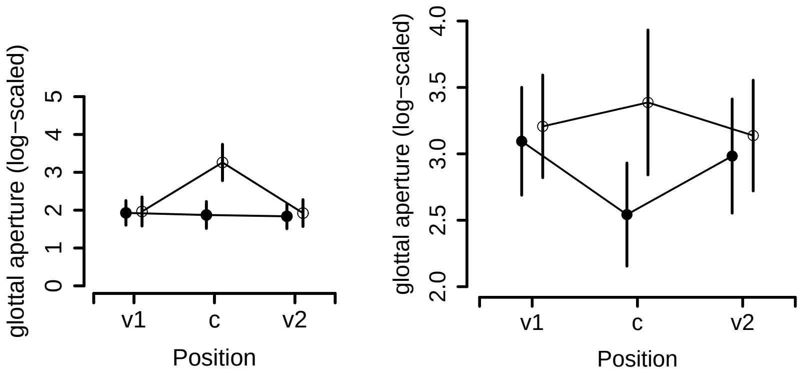


Figure 6.2: Log-scaled glottal aperture plots for spoken (left) and whisperd (right) obstruents (averaged over all speakers, zeros omitted). The points are taken from the vowel preceding the consonant (v1), the consonant itself (c), and the vocalic sequence following the consonant (v2). The filled circles represent voiced tokens, and the open circles represent voiceless tokens. Vertical lines represent one standard deviation above and below each mean.

(rather than with a voiced glottal state, as they are in normal speech). The current data confirm the first half of this prediction—phonologically voiceless consonants are produced with the same glottal aperture in whisper as in normal speech. But the data refute the second half—phonologically voiced consonants are in fact produced with a glottal aperture statistically indistinguishable from the voiced glottal state.

In whispered sequences with voiced obstruents, speakers often adducted the vocal folds from a whisper position on the preceding vowel, to a voiced-like aperture on the obstruents, and back to a whisper position on the following vowel. It is interesting that voiced obstruents and voiced vowels—produced with the same laryngeal configuration in normal speech—have different laryngeal configurations from one another in whispered speech. This suggests that the articulatory specification of voicing on obstruents is different from the articulatory specification of voicing on vowels.

Note, however, that no actual phonetic voicing occurred during the production of phonologically voiced obstruents in whisper, although the glottal aperture is similar to that of phonetically voiced consonants in normal speech. The current data do not allow us to determine why this is. For vocal fold vibration to occur, several conditions need to be met. The vocal folds need to have not only appropriate positioning, but also the correct tension and elasticity (Borden & Harris

1984, p83). In addition, subglottal pressure and airflow across the glottis need to be controlled. Monoson & Zemlin (1984) and Stathopoulos, Hoit, Hixon, Watson & Solomon (1991) report significantly greater airflow in whisper than in normal speech. Stathopoulos et al. (1991) calculate airway resistance from their measures of airflow and oral pressure, and conclude that airway resistance in the larynx is lower in whispered than in normal speech. These results do not tell us exactly what laryngeal and sub-laryngeal articulatory modifications are made; they do, however, point to radically different aerodynamic conditions in whispered speech than in normal speech. These aerodynamic differences, in combination with possible laryngeal articulatory differences not captured by our simple glottal aperture measure, are the likely reason why we do not observe phonetic voicing in phonologically voiced obstruents in whispered speech, although the glottal aperture is the same as it is in normal speech when the segments are voiced.

6.1.2 Perceptual study

In the perceptual test, we discover that listeners can discriminate phonologically voiced from voiceless obstruents in whisper, with 75% accuracy on plosives and 56% accuracy on fricatives (both significantly above chance). Figure 6.3

While some loss of perceptual discrimination is clear, it is not complete. Both classes of sounds—fricatives and stops—are recognized significantly above chance levels (50%). Because of this, the possibility remains that the glottal aperture distinction observed in the visual data is driven by the goal of maintaining a perceptual contrast—an acoustic target, rather than an articulatory target.

Whisper is not, therefore, an effective tool for probing the question of articulatory versus acoustic targets in glottal contrasts. Any tool used to investigate motor control variables must cause a systematic perturbation to some relevant aspect of production. The articulatory measurements show that whispered speech does not detectably perturb the production of voicing contrasts in obstruents (at least with respect to glottal aperture). The perceptual study shows that whispered speech also fails to systematically remove acoustic cues to those contrasts.

The much lower perceptual discriminability of fricatives than plosives (also apparent in normal speech: fricatives at 86% and plosives at 99%) is also of interest. I hypothesize, following the speculations in section 3.2, that acoustic differences

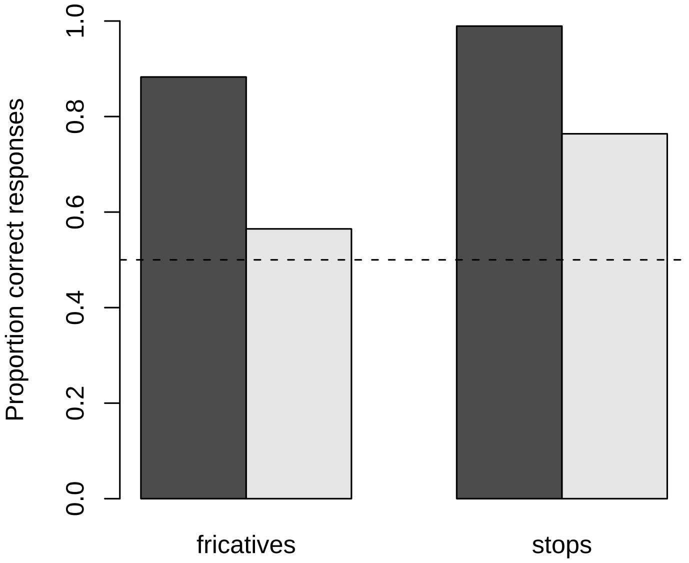


Figure 6.3: Summary of perceptual scores on the voicing contrasts in the endoscopy study, subdivided by manner. Dark bars indicate performance on normal speech; light bars indicate performance on whispered speech.

in the release burst of stops are exploited by listeners to discriminate them in both normal and whispered speech.

I observed that the release bursts of voiceless stops seem to have greater energy than those of voiced stops. Aerodynamically, this is probably due to the greater closure duration of voiceless stops. Greater closure duration would cause greater buildup of pressure posterior to the obstruction, which generates a higher-energy burst upon release of the closure. This would explain why the acoustic difference is preserved in whisper—because the duration difference is the same in whispered as in normal speech. It would also explain why fricatives (which, having incomplete closure, do not cause a buildup of pressure) are not so readily discriminated in normal or whispered speech.

Remember that the glottal aperture differences between phonologically voiced and voiceless consonants are preserved in both whispered stops and whispered fricatives. Statistically, there was no significant difference between glottal apertures on consonants in normal speech and those in whispered speech.

6.1.3 Gestural account

While the results do not allow us to decide between an articulatory and an acoustic account of motor control, they do tell us somethig about what such accounts must look like.

Gestural accounts, as commonly conceived, are captured well by the gestural scores of Browman & Goldstein's Articulatory Phonology. Figure 6.4, for example, illustrates a gestural score for the word "tier" ([tʰiɹ] as spoken by our Canadian English participants).

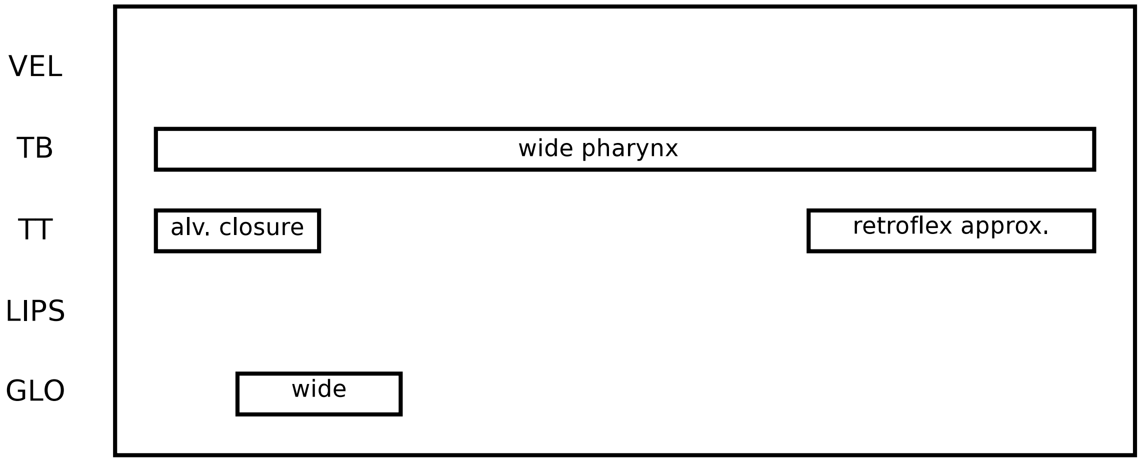


Figure 6.4: Gestural score for "tier". The labels for tiers of the score are VEL=velum, TB=tongue body, TT=tongue tip, LIPS=lips, and GLO=glottis.

Note in particular the glottal widening gesture that overlaps with the tongue tip gesture. One of the questions in this study was whether that gesture would be expressed in whisper, or whether the whisper "posture" would override it. We can now say that the gesture *is* expressed.

Figure 6.5 presents a gestural score for the word "deer" ([tiɹ] or [diɹ]), as it would be given by Browman & Goldstein (1992, and later). Notice that it does not specify a glottal gesture to accompany the alveolar closure.

This underspecification is based on the very sensible idea that, where no change in articulator position is observed, no phonological specification for a gesture should be posited. The current data confirm that there is no change in glottal posture when the [d] in "deer" is spoken normally—it remains closed and vibrating throughout the vowel-consonant-vowel sequence.

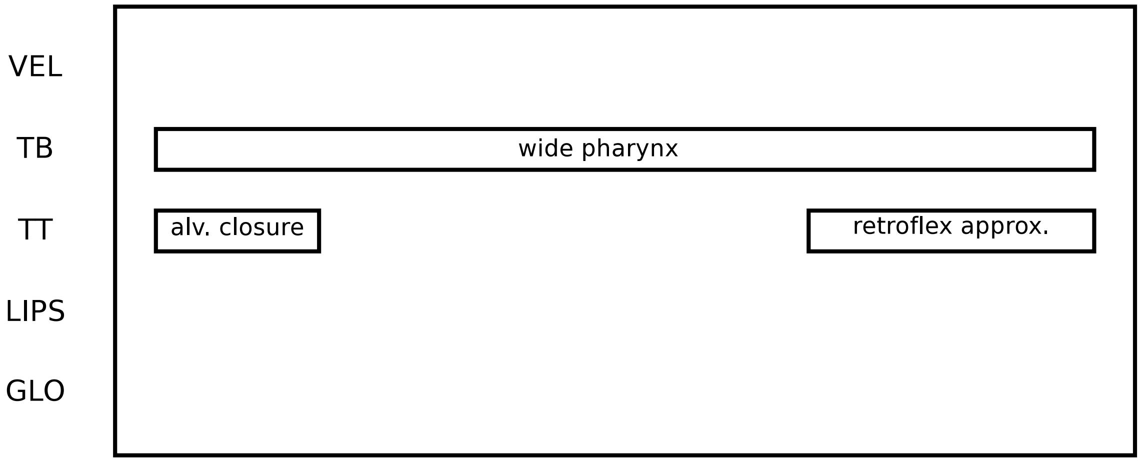


Figure 6.5: Gestural score for "deer", based on Browman & Goldstein (1992). The labels for tiers of the score are VEL=velum, TB=tongue body, TT=tongue tip, LIPS=lips, and GLO=glottis.

However, our observations of whispered speech showed a statistically significant adduction gesture between the phonetically whispered vowels and the [d]. So this suggests that, at least in whispered speech, the gestural score for "deer" should be as depicted in figure 6.6. Note that the exact alignment of the glottal "narrow" gesture cannot be determined from the current results; we can only say that it includes at least the release of the oral closure.

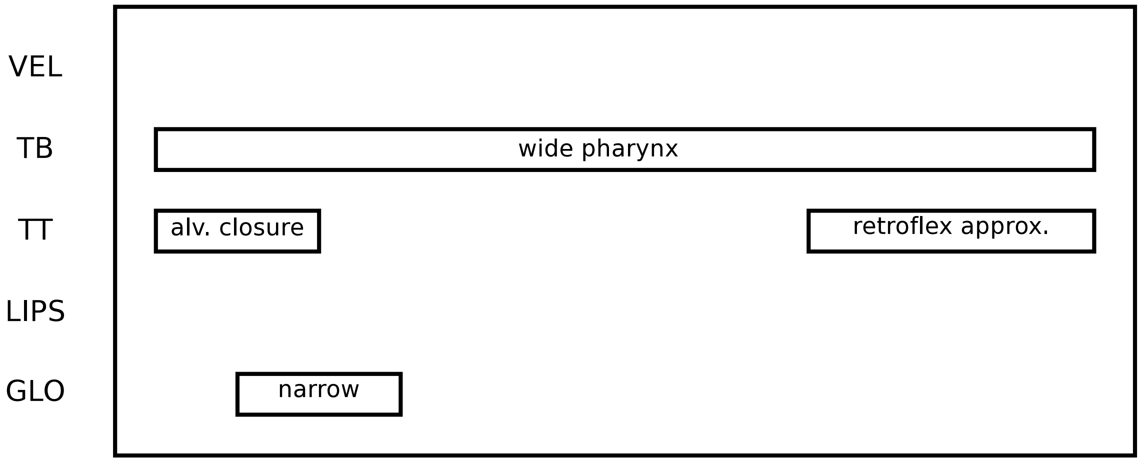


Figure 6.6: Revised gestural score for "deer", given the current results. The labels for tiers of the score are VEL=velum, TB=tongue body, TT=tongue tip, LIPS=lips, and GLO=glottis.

It would seem perverse for this score to be used in whispered speech, and the one from figure 6.5 to be used in normal speech. Parsimony suggests that the new score, with the explicit "narrow glottis" gesture for the [d], is used for both. This means we have a gestural specification ("narrow glottis" on voiced obstruents) which does not translate to a distinct physical gesture in normal speech.

6.1.4 The phonology of glottal states

Even if we do not take a gestural approach to the mental respresentation of phonological structure, the asymmetry of laryngeal behaviour between vowels and phonologically voiced consonants in the current data is instructive.

Phonologically voiceless obstruents have a "voiceless" (fully opened) glottal aperture in both normal and whispered speech, and phonologically voiced obstruents have a "voiced" (fully narrowed) glottal aperture in both normal and whispered speech. On the other hand, vowels have a "voiced" (fully narrowed) glottal aperture in normal speech, and a "whispered" (intermediate between opened and narrowed) glottal aperture in whispered speech. When speakers whisper, they change the glottal aperture of their vowels, but not that of their consonants.

The obvious way to explain this is to invoke the system of phonological contrasts. In English, obstruents are contrastive for phonological voicing but vowels are not. However pronunciations are encoded in the mental lexicon, minimal pairs such as "peer" and "beer" must be distinguished—the former specifying a wide glottis in the onset; the latter specifying a narrow glottis. Vowels, on the other hand, can take a default value. In normal speech, they are voiced; in whispered speech, they are whispered.

One might ask which is primary—do vowels take on whisper phonation because the speaker is adopting a "whispered" articulatory posture; or is the speech whispered because the speaker is using whisper phonation on vowels? The current data cannot empirically rule on this question. However, using the concept of underspecification, I prefer to think that vowels take their glottal state from the underlying posture.

6.2 Contrastive accent

The study in chapter 5 looks at how different parameters are coordinated in a prosodic linguistic variable, to see how motor control of prosody compares to that of the segmental variables that form the bulk of speech motor control research to date.

The prosodic variable I used was relative prominence, varied by eliciting target words in contrastive and non-contrastive contexts. Baseline results showed that, of the five acoustic measures tested, three (f_0 peak, syllable duration, and amplitude peak) were used systematically to signal contrast and two (spectral tilt and vowel quality) were not.

6.2.1 Question manipulation

The question context successfully reduced the contribution of f_0 to the signalling of contrastive context. There was no compensatory increase in the contributions of duration and amplitude; in fact, both duration and amplitude showed *decreased* effects of contrastive context in questions, relative to statements.

Independent parameters systematically varying sympathetically, rather than compensatorily, with the experimentally-manipulated parameter is not a behaviour predicted or explainable under standard accounts of motor control variables.

The answer may lie in one key difference between the linguistic variables examined in previous compensation studies and contrastive emphasis. Previous studies have looked at linguistic variables with fixed targets: lip closure or vowel height in bite-block studies, and formant values in acoustic motor equivalence studies. For these, the linguistic target is the same from utterance to utterance. The contribution to the target that is required from parameter B (say, lip movement) to achieve the target is fully determined by the contribution from parameter A (say, jaw movement). This is schematized in equation 6.1 (where T is the fixed target):

$$A + B = T \tag{6.1}$$

For contrastive accent (and other linguistic variables based on relative prominence), the relationship is different. The combined effect of parameters A, B, and C (say, f_0, duration, and amplitude) must exceed a particular target, which is defined by another point in the utterance or discourse, in which the parameters have values A_0, B_0, and C_0. This is schematized in equation 6.2:

$$A + B + C > A_0 + B_0 + C_0 \tag{6.2}$$

In this case, a disruption in which the contribution of A is fixed at some abnormal level does not fully specify the combined contribution of B and C that is required, because the target is relative rather than fixed.

Also, the studies involving bite block and acoustic motor equivalence involve linguistic variables where it is relatively easy to determine the extent to which A and B contribute to the target, in articulatory or acoustic terms. If the target lip aperture is 0 mm and the current aperture is (say) 6 mm, then a 1 mm vertical movement of the jaw must be accompanied by a 5 mm movement of the lips. The calculations for achieving a particular second-formant frequency may be less straightforward, but they are nevertheless deterministic: a particular tongue posture will require a specific lip protrusion to achieve a specified formant frequency.

For the current study, what does it mean for $A + B + C$ to be more prominent than $A_0 + B_0 + C_0$? A reasonable answer would be that prominence is perceptually determined—it is an empirical question to map the ways in which different acoustic parameters interact in the perception of relative prominence. However, it could be that these interactions are too complex for (or otherwise inaccessible to) the motor control variables used in speech production. In that case, what possible strategies are available to speakers for controlling the acoustic parameters in question? One would be an "every-parameter-for-itself" strategy—the null hypothesis where we would expect no response of one parameter to a perturbation of the others. Clearly this isn't happening in the question manipulation.

Another strategy might involve controlling the different parameters as a bundle: when one goes up, the others go up; when one goes down, the others go down. That is what we see in the question manipulation. This would seem to reduce the dimensionality of the motor control task.

This suggestion is not predicted by previous data and theory on motor control; but the task observed in this study is qualitatively different from previous tasks, as mentioned.

A useful test of this idea of "bundled parameters" might be gained if a non-speech motor task could be identified which shared one or both the qualitative properties of relative prominence identified above: a relative rather than a fixed target, and a complex rather than a simple relationship between the articulatory contributions and the target. Examination of other linguistic tasks sharing these properties (as well as replication of the current results) could also yield insight.

6.2.2 Whisper manipulation

In the whisper manipulation, we saw no compensation in either duration or amplitude for the lack of an f_0 cue. In the absence of the data from the endoscopy chapter and from the question manipulation, this result would lead us to the conclusion that the different parameters are governed by separate control variables.

However, the endoscopy data calls into question the utility of whispered speech in examining control variables. It is possible that the articulatory manipulations that normally generate f_0 contours are present in whispered speech, but are acoustically ineffective (having no glottal vibrations to act on). If this is the case, then the whisper "manipulation" might not be perturbing the speech at all, from the perspective of the control variables. Support for this possibility comes from the question manipulation. There, we know that f_0 is perturbed, and we see the sympathetic response of duration and amplitude. The lack of such a response in whisper means either that the whisper manipulation failed to actually perturb the articulation of f_0, or that whispered speech is controlled in a systematically different manner than normal speech. The former possibility is much more likely than the latter in the light of the endoscopy findings.

We might also remember the general conclusion of motor control theory—that "normal speech production programming is indeed 'compensatory' .... The differences between compensatory and normal articulation do not reside in the choice of different encoding strategies, but rather have to do with extreme versus non-extreme articulatory parameter values." (Lindblom et al. 1979, p159) It seems more likely that the whisper manipulation failed to disrupt f_0 articulations (as it failed to disrupt "voicing" articulations in the endoscopy study) than that we are seeing two completely different modes of speech motor control at work.

6.3 Summary

This dissertation presents two studies exploring the nature of motor control variables in speech production.

The first experiment, looking at glottal aperture in whispered voicing contrasts, reveals that the same glottal aperture differences seen in normal speech between phonologically voiced and voiceless obstruents are also seen in whispered speech; this difference is shown to be perceptible to listeners, though the acoustic property signalling the difference is not identified. Phonological voicing in vowels and phonological voicing in obstruents are shown in whispered speech to operate in qualitatively different ways.

The second experiment, investigating the acoustic contributors to contrastive emphasis (frequency, duration, and amplitude), finds no evidence of compensation; however, an unexpected “sympathetic attenuation” is observed in the question manipulation which is not predicted by the standard control variable paradigm. This suggests that motor control of prosodic variables may be qualitatively different from motor control of segmental variables, though more research is required to determine the generality of this behaviour.

Appendix A

Glossary

This work requires the ability to distinguish between certain meanings for which commonly-agreed specific terms do not yet emerge from the literature at large. Also included are some terms that may be unfamiliar to phoneticians who have not worked with endoscopic data. This glossary specifies the meanings used in this work.

abduct (v) to place farther apart (also *abducted*, *abduction*)—used in this work to refer to the opening of the vocal folds.

adduct (v) to place closer together (also *adducted*, *adduction*)—used in this work to refer to the closing of the vocal folds.

accent (n) a physical manifestation of semantic or syntactic prominence on a syllable. To be distinguished from *stress* and *focus*, which are abstract properties of syllables and phrases, respectively, and from *prominence*, which is a perceptual property of various levels of constituents.

contrastive accent (n) a pitch accent produced in a contrastive context

control variable (n) see *motor control variable*

coordinative structure (n) the set of muscles and articulators which act together to achieve a particular target specified by a single motor control variable.

cuneiform tubercles (n) two cartilaginous structures that rest on top of the arytenoid cartilages in the larynx. The cuneiform tubercles are visible in endoscopic video recordings as bumps in the anterior/inferior ends of the aryepiglottal folds.

goal (n) the consciously intended end for which an action is undertaken

motor control variable (n) the final organisational specification of an action in the brain before the motor commands are sent to the muscles. Also referred to simply as "control variable". Control variables can govern multiple muscles and even multiple articulators—for example, a control variable specifying hand position governs (at least) extension, flexion, and rotation in the shoulder, elbow, and wrist.

normal speech (n) a sequence of speech in which at least some segments are produced with vocal fold vibration. This is the normal speech mode of most healthy speakers.

phonologically voiced (adj) of a segment or sequence of segments, belonging to an abstract category in a language's phonology which is characterized by the presence of phonetic voicing

phonologically voiceless (adj) of a segment or sequence of segments, belonging to an abstract category in a language's phonology which is characterized by the absence of phonetic voicing

prominence (n) a generic term, often used to indicate the perceptual salience of one part of an utterance relative to another part. Prominence can reflect any of a large number of possible semantic and lexical relations between items. Some examples include unstressed and stressed syllables in a word, function and content words in a phrase, and given and new information in an utterance.

voiced (adj) of a segment or sequence of segments, produced with a glottal state in which the vocal folds are regularly vibrating

voiceless (adj) of a segment or sequence of segments, produced with a glottal state in which there is no vocal fold vibration and the vocal folds are sufficiently abducted to prevent turbulent excitation of the transglottal airstream (for voicelessness in a phonological sense, see "phonologically voiceless")

whisper (n) a glottal state in which the vocal folds are not vibrating, but are adducted so that turbulent noise is generated as air passes through the larynx

whispered (adj) of a segment or sequence of segments, produced with whisper

whispered speech (n) a sequence of speech in which there is no vocal fold vibration, and in which the primary glottal excitation is whispered. (Note that other glottal states may occur in whispered speech, so long as they do not involve vocal fold vibration.)

Appendix B

Endoscopy study: Sentences read

Following is the full set of sentences elicited in the recordings for the endoscopy study described in chapter 1. Only the basic voicing contrasts from the sentences in table B.1 were analysed for this dissertation.

B.1 Voicing contrasts

For each speaker, the basic voicing contrasts (table B.1) were elicited first, followed by the emphasized targets (table B.2), followed by the unemphasized targets (table B.3).

1	Say PEER again.
2	Say TIER again.
3	Say FEAR again.
4	Say SEAL again.
5	Say BEER again.
6	Say DEAR again.
7	Say VEER again.
8	Say ZEAL again.

Table B.1: Orthographic presentation form of sentences eliciting voicing contrast

B.2 Pitch gestures

To elicit different pitch gestures, a contrastive emphasis situation was constructed. In one condition (emphasised, higher pitch), the frame sentence was “Say *x* again, not *y* again” where pairs of target words were taken from table B.1. The other condition (non-emphasised, lower pitch) used the frame sentence “*Say* x again, don't *write* x again", where the emphatic contrast is on the word immediately preceding the target word.

In order to control for segmental environment, only the first target word was measured in each case. Eight sentences with emphatic targets were used—one with each of the eight words in initial position (x) and its minimal pair in the other position (y) (table B.2). For a different pitch pattern, eight sentences with non-emphatic targets were used—one for each of the eight words (table B.3).

1	Say PEER again, not BEER again.
2	Say BEER again, not PEER again.
3	Say TIER again, not DEAR again.
4	Say DEAR again, not TIER again.
5	Say FEAR again, not VEER again.
6	Say VEER again, not FEAR again.
7	Say SEAL again, not ZEAL again.
8	Say ZEAL again, not SEAL again.

Table B.2: Orthographic presentation form of sentences in the pitch study—target words emphasised.

1	SAY peer again, don't WRITE peer again.
2	SAY beer again, don't WRITE beer again.
3	SAY tier again, don't WRITE tier again.
4	SAY dear again, don't WRITE dear again.
5	SAY fear again, don't WRITE fear again.
6	SAY veer again, don't WRITE veer again.
7	SAY seal again, don't WRITE seal again.
8	SAY zeal again, don't WRITE zeal again.

Table B.3: Orthographic presentation form of sentences in pitch study—target words not emphasised.

APPENDIX C

Prosody study: Sentences read

The following tables present all of the sentences used in the prosody experiment (chapter 5). They are organized first by target word, and second by prosodic condition.

The first ten tables (tables C.1 through C.10) divide the sentences by target word—each table gives the prompts for a given target word in all four prosodic conditions.

The remaining four tables (tables table:statement-non-contrastive-sentences through table:qu contrastive-sentences) present the same sentences by the prosodic conditions. It

nc	st	The passengers will feel safe.	Their seatbelts will be fastened.
c	st	Their seatbelts won't be released.	Their seatbelts will be fastened.
nc	qu	The passengers will feel safe.	Will their seatbelts be fastened?
c	qu	Their seatbelts won't be released.	Will their seatbelts be fastened?

Table C.1: Sentences elicited in prosody experiment for the target word "fastened". Prosodic conditions indicated: nc=non-contrastive, c=contrastive, st=statement, qu=question.

nc	st	The race will go well.	She can cycle fastest.
c	st	She can't cycle longest.	She can cycle fastest.
nc	qu	The race will go well.	Can she cycle fastest?
c	qu	She can't cycle longest.	Can she cycle fastest?

Table C.2: Sentences elicited in prosody experiment for the target word "fastest". Prosodic conditions indicated: nc=non-contrastive, c=contrastive, st=statement, qu=question.

nc	st	They won't be in tomorrow.	They're golfing with their father.
c	st	They're not golfing with their mother.	They're golfing with their father.
nc	qu	They won't be in tomorrow.	Are they golfing with their father?
c	qu	They're not golfing with their mother.	Are they golfing with their father?

Table C.3: Sentences elicited in prosody experiment for the target word "father". Prosodic conditions indicated: nc=non-contrastive, c=contrastive, st=statement, qu=question.

nc	st	The plane went down nearby.	They're searching the marshes.
c	st	They're not searching the forest.	They're searching the marshes.
nc	qu	The plane went down nearby.	Are they searching the marshes?
c	qu	They're not searching the forest.	Are they searching the marshes?

Table C.4: Sentences elicited in prosody experiment for the target word "marshes". Prosodic conditions indicated: nc=non-contrastive, c=contrastive, st=statement, qu=question.

nc	st	That show is my favorite.	It's on in the morning.
c	st	It's not on in the evening.	It's on in the morning.
nc	qu	That show is my favorite.	Is it on in the morning?
c	qu	It's not on in the evening.	Is it on in the morning?

Table C.5: Sentences elicited in prosody experiment for the target word "morning". Prosodic conditions indicated: nc=non-contrastive, c=contrastive, st=statement, qu=question.

nc	st	He found a small error.	He checked his main sources.
c	st	He didn't check his main data.	He checked his main sources.
nc	qu	He found a small error.	Did he check his main sources?
c	qu	He didn't check his main data.	Did he check his main sources?

Table C.6: Sentences elicited in prosody experiment for the target word "sources". Prosodic conditions indicated: nc=non-contrastive, c=contrastive, st=statement, qu=question.

nc	st	They're away today.	They're going surfing.
c	st	They're not going swimming.	They're going surfing.
nc	qu	They're away today.	Are they going surfing?
c	qu	They're not going swimming.	Are they going surfing?

Table C.7: Sentences elicited in prosody experiment for the target word "surfing". Prosodic conditions indicated: nc=non-contrastive, c=contrastive, st=statement, qu=question.

nc	st	They had a light dinner.	They ordered some sushi.
c	st	They didn’t order some pizza.	They ordered some sushi.
nc	qu	They had a light dinner.	Did they order some sushi?
c	qu	They didn’t order some pizza.	Did they order some sushi?

Table C.8: Sentences elicited in prosody experiment for the target word “sushi”. Prosodic conditions indicated: nc=non-contrastive, c=contrastive, st=statement, qu=question.

nc	st	She’s still in her office.	She’s writing a thesis.
c	st	She’s not writing a novel.	She’s writing a thesis.
nc	qu	She’s still in her office.	Is she writing a thesis?
c	qu	She’s not writing a novel.	Is she writing a thesis?

Table C.9: Sentences elicited in prosody experiment for the target word “thesis”. Prosodic conditions indicated: nc=non-contrastive, c=contrastive, st=statement, qu=question.

nc	st	He ran for hours.	He’s very thirsty.
c	st	He’s not very hungry.	He’s very thirsty.
nc	qu	He ran for hours.	Is he very thirsty?
c	qu	He’s not very hungry.	Is he very thirsty?

Table C.10: Sentences elicited in prosody experiment for the target word “thirsty”. Prosodic conditions indicated: nc=non-contrastive, c=contrastive, st=statement, qu=question.

was in these sets, randomized separately in five repetitions, that the items were presented to speakers.

They had a light dinner.	They ordered some sushi.
She's still in her office.	She's writing a thesis.
He ran for hours.	He's very thirsty.
That show is my favorite.	It's on in the morning.
They're away today.	They're going surfing.
The race will go well.	She can cycle fastest.
The passengers will feel safe.	Their seatbelts will be fastened.
They won't be in tomorrow.	They're golfing with their father.
He found a small error.	He checked his main sources.
The plane went down nearby.	They're searching the marshes.

Table C.11: Sentences elicited in prosody experiment for statement intonation and non-contrastive accent.

Their seatbelts won't be released.	Their seatbelts will be fastened.
She can't cycle longest.	She can cycle fastest.
They're not golfing with their mother.	They're golfing with their father.
They're not searching the forest.	They're searching the marshes.
It's not on in the evening.	It's on in the morning.
He didn't check his main data.	He checked his main sources.
They're not going swimming.	They're going surfing.
They didn't order some pizza.	They ordered some sushi.
She's not writing a novel.	She's writing a thesis.
He's not very hungry.	He's very thirsty.

Table C.12: Sentences elicited in prosody experiment for statement intonation and contrastive accent.

The passengers will feel safe.	Will their seatbelts be fastened?
The race will go well.	Can she cycle fastest?
They won't be in tomorrow.	Are they golfing with their father?
The plane went down nearby.	Are they searching the marshes?
That show is my favorite.	Is it on in the morning?
He found a small error.	Did he check his main sources?
They're away today.	Are they going surfing?
They had a light dinner.	Did they order some sushi?
She's still in her office.	Is she writing a thesis?
He ran for hours.	Is he very thirsty?

Table C.13: Sentences elicited in prosody experiment for question intonation and non-contrastive accent.

Their seatbelts won't be released.	Will their seatbelts be fastened?
She can't cycle longest.	Can she cycle fastest?
They're not golfing with their mother.	Are they golfing with their father?
They're not searching the forest.	Are they searching the marshes?
It's not on in the evening.	Is it on in the morning?
He didn't check his main data.	Did he check his main sources?
They're not going swimming.	Are they going surfing?
They didn't order some pizza.	Did they order some sushi?
She's not writing a novel.	Is she writing a thesis?
He's not very hungry.	Is he very thirsty?

Table C.14: Sentences elicited in prosody experiment for question intonation and contrastive accent.

References

Abercrombie, David (1967), *Elements of General Phonetics*, Edinburgh University Press, Aberdeen.

Adobe (2001), 'Premiere', Computer program (version 6.0).
URL: *http://www.adobe.com/*

Anastaplo, Sara and Michael P. Karnell (1988), 'Synchronized videostroboscopic and electroglottographic examination of glottal opening', *Journal of the Acoustical Society of America* **83**(5), 1883–1890.

Aylett, Matthew and Alice Turk (2006), 'Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei', *Journal of the Acoustical Society of America* **119**(5), 3048–3058.

Baayen, R. H. (2008), *Analyzing Linguistic Data: A practical introduction to statistics*, Cambridge University Press.

Baer, T., A. Lofquist, and N. McGarr (1983), 'Laryngeal vibrations: A comparison between high-speed filming and glottographic techniques', *Journal of the Acoustical Society of America* **73**(4), 1304–1308.

Bates, Douglas (2005), 'Using the lme4 package', *R News* **5**(1), 27–30.

Baum, Shari R. (1988), 'Acoustic analysis of compensatory articulation in children', *Journal of the Acoustical Society of America* **84**(5), 1662–1668.

Baum, Shari R. (1999), 'Compensations for jaw fixation by aphasic patients under conditions of increased articulatory demands: a follow-up study', *Aphasiology* **13**(7), 513–527.

Baum, Shari R. and David H. McFarland (1997), 'The development of speech adaptation to an artificial palate', *Journal of the Acoustical Society of America* **102**(4), 2353–2359.

Baum, Shari R., Jean A. Kim, and William F. Katz (1997), 'Compensation for jaw fixation by aphasic patients', *Brain and Language* **56**, 354–376.

Benguerel, André-Pierre and Tej K. Bhatia (1980), 'Hindi stop consonants: an acoustic and fiberscopic study', *Phonetica* **37**, 134–148.

Benguerel, André-Pierre, Hajime Hirose, M. Sawashima, and T. Ushijima (1978), 'Laryngeal control in French stop production: a fiberscopic, acoustic and electromyographic study', *Folia Phoniatrica* **30**, 175–198.

Berkovits, Rochele (1984), 'Duration and fundamental frequency in sentence-final intonation', *Journal of Phonetics* **12**, 255–265.

Boersma, Paul and David Weenink (2005), 'Praat: doing phonetics by computer'. [computer program], retrieved 2005.
URL: *http://www.praat.org/*

Borden, Gloria J. and Katherine S. Harris (1984), *Speech Science Primer*, 2nd edn, Williams and Wilkins, Baltimore, USA.

Botinis, Antonis, Georgios Kouroupetroglou, and George Carayiannis (eds.) (1997), *Intonation: Theory, Models and Applications*, European Speech Communication Association, S. Athanasopoulos, Athens, Greece.

Braun, Bettina (2004), Production and Perception of Contrastive and Non-Contrastive Themes in German, PhD thesis, Universität des Saarlandes.

Browman, Catherine P. and Louis Goldstein (1992), 'Articulatory phonology: An overview', *Phonetica* **49**, 155–180.

Browman, Catherine P. and Louis M. Goldstein (2000), 'Competing constraints on intergestural coordination and self-organization of phonological structures', *Les Cahiers de l'ICP, Bulletin de la Communication Parlée* **5**, 25–34.

Campbell, Nick and Mary E. Beckman (1997), Stress, prominence, and spectral tilt, *in* Botinis, Kouroupetroglou & Carayiannis (1997), pp. 67–70.

Catford, J. C. (1964), Phonation types: The classification of some laryngeal components of speech production, *in* David Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott, and J. L. M. Trim (Eds.), 'In Honour of Daniel Jones', Longmans, Green and Co. Ltd., London, pp. 26–37.

Catford, J. C. (1977), *Fundamental Problems in Phonetics*, Edinburgh University Press, Edinburgh, UK.

Cooper, William E., Stephen J. Eady, and Pamela R. Mueller (1985), 'Acoustical aspects of contrastive stress in question-answer contexts', *Journal of the Acoustical Society of America* **77**(6), 2142–2156.

Crystal, Thomas H. and Arthur S. House (1988), 'Segmental durations in connected-speech signals: current results', *Journal of the Acoustical Society of America* **83**(4), 1553–1573.

Dailey, Seth H., James B. Kobler, Robert E. Hillman, Kittisard Tangrom, Ekawudh Thananart, Marcelo Mauri, and Steven M. Zeitels (2005), 'Endoscopic measurement of vocal fold movement during adduction and abduction', *The Laryngoscope* **115**(1), 178–183.

Dannenbring, Gary L. (1980), 'Perceptual discrimination of whispered phoneme pairs', *Perceptual and Motor Skills* **51**, 979–985.

E-Studio (2003), 'E-studio perceptual experiment software', Computer program (version 1.1.4.15).
URL: *http://www.pstnet.com*

Eady, Stephen J. and William E. Cooper (1986), 'Speech intonation and focus location in matched statements and questions', *Journal of the Acoustical Society of America* **80**(2), 402–415.

Edmondson, Jerold A. and John H. Esling (2006), 'The valves of the throat and their functioning in tone, vocal register and stress: laryngoscopic case studies', *Phonology* **23**, 157–191.

Edwards, Jan (1992), 'Compensatory speech motor abilities in normal and phonologically disordered children', *Journal of Phonetics* **20**, 189–207.

Esling, John H. (2002), The laryngeal sphincter as an articulator: How register and phonation interact with vowel quality and tone, *in* 'WECOL 2002', WECOL, UBC, pp. 68–86.

Esling, John H. and Jimmy G. Harris (2003), An expanded taxonomy of states of the glottis, *in* 'Proceedings of the 15th International Congress of Phonetic Sciences', Vol. 1, UAB, Barcelona, pp. 1049–1052.

Esling, John H. and Jimmy G. Harris (2005), States of the glottis: An articulatory phonetic model based on laryngoscopic observations, *in* Hardcastle & Beck (2005), chapter 14, pp. 347–383.

Fant, Gunnar (1960), *Acoustic Theory of Speech Production*, Mouton & Co, The Hague.

Flanagan, James L. (1955), 'Difference limen for the intensity of a vowel sound', *Journal of the Acoustical Society of America* **27**, 1223–1225.

Flash, Tamar and Neville Hogan (1985), 'The coordination of arm movements: An experimentally confirmed mathematical model', *The Journal of Neuroscience* **5**(7), 1688–1703.

Flege, J, S Fletcher, and A Homiedan (1988), 'Compensating for a bite block in /s/ and /t/ production: Palatographic, acoustic, and perceptual data', *Journal of the Acoustical Society of America* **83**, 212–228.

Folkins, John W. and James H. Abbs (1975), 'Lip and jaw motor control during speech: responses to resistive loading of the jaw', *Journal of Speech and Hearing Research* **18**, 207–220.

Fowler, Carol A. and Michael T. Turvey (1981), 'Immediate compensation in bite-block speech', *Phonetica* **37**, 306–326.

Fry, D. B. (1955), 'Duration and intensity as physical correlates of linguistic stress', *Journal of the Acoustical Society of America* **27**(4), 765–768.

Fry, D. B. (1958), 'Experiments in the perception of stress', *Language and Speech* **1**(2), 126–152.

Fujimura, O., Thomas Baer, and Seiji Niimi (1979), 'A stereo-fiberscope with a magnetic interlens bridge for laryngeal observation', *Journal of the Acoustical Society of America* **65**(2), 478–480.

Fujisaki, H., K. Nakamura, and T. Imoto (1975), Auditory perception of duration of speech and non-speech stimuli, *in* G. Fant and M.A.A. Tatham (Eds.), 'Auditory Analysis and Perception of Speech', Academic, London, pp. 197–219.

Fulop, Sean A., Ethelbert Kari, and Peter Ladefoged (1998), 'An acoustic study of the tongue root contrast in Degema vowels', *Phonetica* **55**, 80–98.

Gao, Man (2002), Tones in whispered Chinese: Articulatory features and perceptual cues, Master of arts, University of Victoria.

Gay, Thomas, Björn Lindblom, and James Lubker (1981), 'Production of bite-block vowels: Acoustic equivalence by selective compensation', *Journal of the Acoustical Society of America* **69**(3), 802–810.

GIMP (2007), 'GNU Image Manipulation Program', Computer program (version 2.6.2). Licensed under the GNU General Public License.
URL: *http://www.gimp.org*

Grabe, Esther, Brechtje Post, Francis Nolan, and Kimberley Farrar (2000), 'Pitch accent realization in four varieties of British English', *Journal of Phonetics* **28**, 161–185.

Gracco, Vincent L. and Anders Löfqvist (1994), 'Speech motor coordination and control: Evidence from lip, jaw, and laryngeal movements', *The Journal of Neuroscience* **14**(11), 6585–6597.

Guenther, Frank H., Cary Y. Espy-Wilson, Suzanne E. Boyce, Melanie L. Matthies, Majid Zandipour, and Joseph S. Perkell (1999), 'Articulatory trade-offs reduce acoustic variability during American English /r/ production', *Journal of the Acoustical Society of America* **105**, 2854–2865.

Guion, Susan G., Mark W. Post, and Doris L. Payne (2004), 'Phonetic correlates of tongue root vowel contrasts in Maa', *Journal of Phonetics* **32**, 517–542.

Gussenhoven, C. and A. C. M. Rietveld (1988), 'Fundamental frequency declination in Dutch: testing three hypotheses', *Journal of Phonetics* **16**, 355–369.

Hamlet, Sandra L. and Maureen Stone (1976), 'Compensatory vowel characteristics resulting from the presence of different types of experimental prostheses', *Journal of Phonetics* **4**, 199–218.

Hamlet, Sandra L. and Maureen Stone (1978), 'Compensatory alveolar consonant production induced by wearing a dental prosthesis', *Journal of Phonetics* **6**, 227–248.

Hanson, Helen M. (1997), 'Glottal characteristics of female speakers: Acoustic correlates', *Journal of the Acoustical Society of America* **101**(1), 466–481.

Hanson, Helen M. and Erika S. Chuang (1999), 'Glottal characteristics of male speakers: Acoustic correlates and comparison with female data', *Journal of the Acoustical Society of America* **106**(2), 1064–1077.

Hanson, Helen M., Kenneth N. Stevens, Hong-Kwang Jeff Kuo, Marilyn Y. Chen, and Janet Slifka (2001), 'Towards models of phonation', *Journal of Phonetics* **29**, 451–480.

Hardcastle, William J. and Janet Mackenzie Beck (eds.) (2005), *A Figure of Speech: A Festschrift for John Laver*, Lawrence Erlbaum Associates, Inc., New Jersey.

Hayden, E. H. and Y. Koike (1972), 'A data processing scheme for frame by frame film analysis', *Folia Phoniatrica* **24**, 169–181.

Heldner, Mattias (2001), Spectral emphasis as an additional source of information in accent detection, *in* 'Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding', ISCA, Red Bank, NJ, pp. 57–60.

Henrich, Nathalie, Christophe d'Alessandro, Boris Doval, and Michèle Catellengo (2004), 'On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation', *Journal of the Acoustical Society of America* **115**(3), 1321–1332.

Hess, Markus M. and Michael Ludwigs (2000), 'Strobophotoglottographic transillumination as a method for the analysis of vocal fold vibration patterns', *Journal of Voice* **14**(2), 255–271.

Higashikawa, Masahiko and Fred D. Minifie (1999), 'Acoustical-perceptual correlates of "whisper pitch" in synthetically generated vowels', *Journal of Speech, Language, and Hearing Research* **42**, 583–591.

Higashikawa, Masahiko, Jordan R. Green, Christopher A. Moore, and Fred D. Minifie (2003), 'Lip kinematics for /p/ and/b/ production during whispered and voiced speech', *Folia Phoniatrica et Logopaedica* **55**, 17–27.

Hirose, Hajime and Tatsujiro Ushijima (1978), 'Laryngeal control for voicing distinction in Japanese consonant production', *Phonetica* **35**, 1–10.

Hirose, Hajime, C. Y. Lee, and Tatsujiro Ushijima (1974), 'Laryngeal control in Korean stop production', *Journal of Phonetics* **2**, 145–152.

Hirst, Daniel and Robert Espesser (1993), 'Automatic modelling of fundamental frequency using a quadratic spline function', *Travaux de l'Institut de Phonétique d'Aix* **15**, 75–85.

Hodson, David (2007), 'Wideangle distortion filter'. Licensed under the GNU General Public License.
URL: *http://members.ozemail.com.au/hodsond/wideangle.html*

Iwata, Ray, Masayuki Sawashima, Hajime Hirose, and Seiji Niimi (1979), 'Laryngeal adjustments of Fukienese stops', *Ann. Bull. RILP* **13**, 61–81.

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

Kagaya, Ryohei (1974), 'A fiberscopic and acoustic study of the Korean stops, affricates and fricatives', *Journal of Phonetics* **2**, 161–180.

Kallail, Ken J. and Floyd W. Emanuel (1984*a*), 'An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects', *Journal of Phonetics* **12**(2), 175–186.

Kallail, Ken J. and Floyd W. Emanuel (1984*b*), 'Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects', *Journal of Speech and Hearing Research* **27**, 245–251.

Kelso, J. A. Scott and Betty Tuller (1983), "'Compensatory articulation" under conditions of reduced afferent information: a dynamic formulation', *Journal of Speech and Hearing Research* **26**, 217–224.

Kiver, Milton S. (1964), *Color Television Fundamentals*, 2nd edn, McGraw-Hill, New York.

Kochanski, G, E Grabe, John Coleman, and B Rosner (2005), 'Loudness predicts prominence: fundamental frequency lends little', *Journal of the Acoustical Society of America* **118**(2), 1038–1054.

Krippendorff, Klaus (1980), *Content Analysis: An Introduction to Its methodology*, Vol. 5 of *The Sage CommText Series*, Sage Publications.

Ladd, D. Robert, Jo Verhoeven, and Karen Jacobs (1994), 'Influence of adjacent pitch accents on each other's perceived prominence: two contradictory effects', *Journal of Phonetics* **22**, 87–99.

Ladefoged, Peter (2005), Speculations on the control of speech, *in* Hardcastle & Beck (2005), chapter 1, pp. 3–21.

Laver, John (1994), *Principles of Phonetics*, Cambridge University Press, Cambridge, UK.

Lee, Avery (2005), 'VirtualDub video processing software', Computer program (build 23604). Licensed under the GNU General Public License.
URL: *http://www.virtualdub.org/*

Lindblom, Björn, James Lubker, and Thomas Gay (1979), 'Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation', *Journal of Phonetics* **7**, 141–161.

Lisker, Leigh, Arthur S. Abramson, Franklin S. Cooper, and Malcolm H. Schvey (1969), 'Transillumination of the larynx in running speech', *Journal of the Acoustical Society of America* **45**(6), 1544–1546.

Liu, Fang and Yi Xu (2007), Question intonation as affected by word stress and focus in English, *in* 'ICPhS Proceedings 2007', International Congress of Phonetic Sciences.

Lockhart, Daniel B and Lena H Ting (2007), 'Optimal sensorimotor transformations for balance', *Nature Neuroscience* **10**(10), 1329–1336.

Lubker, James F. (1979), 'The reorganization times of bite-block vowels', *Phonetica* **36**, 273–293.

MacNeilage, Peter F. (ed.) (1983), *The Production of Speech*, Springer-Verlag, New York.

Malmberg, Bertil (1963), *Phonetics*, Dover Publications, Inc., New York.

McAuliffe, Megan J., Emily Lin, Michael P. Robb, and Bruce E. Murdoch (2008), 'Influence of a standard electropalatography artificial palate upon articulation', *Folia Phoniatrica et Logopaedica* **60**, 45–53.

McFarland, David H. and Shari R. Baum (1995), 'Incomplete compensation to articulatory perturbation', *Journal of the Acoustical Society of America* **97**(3), 1865–1873.

Mean (2008), 'Avidemux video editor', Computer program (version 2.4.3). Licensed under the GNU General Public License.
URL: *http://avidemux.sourceforge.net/*

Meyer-Eppler, Werner (1957), 'Realization of prosodic features in whispered speech', *Journal of the Acoustical Society of America* **29**(1), 104–106.

Mills, Timothy (2003), Cues to voicing contrasts in whispered Scottish obstruents, Master of science, Edinburgh University.

Monoson, Patricia and Willard R. Zemlin (1984), 'Quantitative study of whisper', *Folia Phoniatrica et Logopaedica* **36**, 53–65.

Munro, Murray J. (1990), 'Perception of 'voicing' in whispered stops', *Phonetica* **47**, 173–181.

Neter, John, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman (1996), *Applied Linear Statistical Models*, 4th edn, Irwin.

Ní Chasaide, Ailbhe and Christer Gobl (1997), Voice source variation, *in* 'The Handbook of Phonetic Sciences', Blackwell Publishers.

Nicholson, Hannele and Andreas Teig (2003), How to tell beans from farmers: cues to the perception of pitch accent in whispered Norwegian, *in* 'Proceedings of the 19th Scandinavian Conference of Linguistics', Vol. 31 of *University of Tromsø working papers in language and linguistics*, Tromsø.

Oller, D. Kimbrough and Peter F. MacNeilage (1983), Development of speech production: Perspectives from natural and perturbed speech, *in* MacNeilage (1983), chapter 5.

Pell, Mark D. (2001), 'Influence of emotion and focus location on prosody in matched statements and questions', *Journal of the Acoustical Society of America* **109**(4), 1668–1680.

Perkell, Joseph S., Frank H Guenther, Harlan Lane, Melanie L. Matthies, Pascal Perrier, Jennell Vick, Reiner Wilhelms-Tricarico, and Majid Zandipour (2000), 'A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss', *Journal of Phonetics* **28**, 233–272.

Perkell, Joseph S., Melanie L. Matthies, Mario A. Svirsky, and Michael A. Jordan (1993), 'Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot "motor equivalence" study', *Journal of the Acoustical Society of America* **93**(5), 2948–2961.

Perkins, William, Joanna Rudas, Linda Johnson, and Jody Bell (1976), 'Stuttering: discoordination of phonation with articulation and respiration', *Journal of Speech and Hearing Research* **19**, 509–522.

Pike, Kenneth L (1943), *Phonetics: A critical analysis of phonetic theory and a technic for the practical description of sounds*, University of Michigan Press, Michigan, USA.

Portele, Th. and B. Heuft (1997), 'Towards a prominence-based synthesis system', *Speech Communication* **21**, 61–72.

Quené, Hugo (2007), 'On the just noticeable difference for tempo in speech', *Journal of Phonetics* **35**, 353–362.

Quené, Hugo and Huub van den Bergh (2008), 'Examples of mixed-effects model with crossed random effects and with binomial data', *Journal of Memory and Language* **59**, 413–425.

R (2008), 'R: A language and environment for statistical computing', Computer program (version 2.6.2). Licensed under the GNU General Public License.
URL: *http://www.r-project.org*

Remijsen, Bert (2001), Word-prosodic systems of Raja Ampat languages, PhD thesis, Universiteit Leiden Centre for Linguistics.

Saltzman, Elliot and J. A. Scott Kelso (1987), 'Skilled actions: A task-dynamic approach', *Psychological Review* **94**(1), 84–106.

Sawashima, M. and H. Hirose (1968), 'New laryngoscopic technique by use of fiberoptics', *Journal of the Acoustical Society of America* **43**(1), 168–169.

Sawashima, M. and H. S. Park (1979), 'Laryngeal adjustments for syllable final stops in Korean: Some preliminary results of fiberoptic observation', *Ann. Bull. RILP* **13**, 83–89.

Sawashima, Masayuki (1979), 'Laryngeal control for voicing distinctions: A review of recent works', *Ann. Bull. RILP* **13**, 23–26.

Shiller, Douglas M., David J. Ostry, Paul L. Gribble, and Rafael Laboissière (2001), 'Compensation for the effects of head acceleration on jaw movement in speech', *The Journal of Neuroscience* **21**(16), 6447–6456.

Sluijter, Agaath M. C. and Vincent J. van Heuven (1996), 'Spectral balance as an acoustic correlate of linguistic stress', *Journal of the Acoustical Society of America* **100**(4), 2471–2485.

Smith, Bruce L. and Ann McLean-Muse (1987), 'Effects of rate and bite block manipulation on kinematic characteristics of children's speech', *Journal of the Acoustical Society of America* **81**(3), 747–754.

Stathopoulos, Elaine T., Jeannette D. Hoit, Thomas J. Hixon, Peter J. Watson, and Nancy Pearl Solomon (1991), 'Respiratory and laryngeal function during whispering', *Journal of Speech and Hearing Research* **34**, 761–767.

Stevens, Hanna E. and Robert E. Wickesberg (2002), 'Representation of whispered word-final stop consonants in the auditory nerve', *Hearing Research* **173**, 119–133.

Stevens, Kenneth N. (1998), *Acoustic Phonetics*, number 30 *in* 'Current Studies in Linguistics', The MIT Press, Cambridge, Massachusetts.

Sweet, Henry (1877), *A Handbook of Phonetics including a Popular Exposition of the Principles of Spelling Reform*, McGrath, Maryland, USA. reprinted in 1970.

Sweet, Henry (1906), *A Primer of Phonetics*, 3 edn, Clarendon Press, Oxford.

Tanabe, Masahiro, Kazutomo Kitajima, Wilbur J. Gould, and Anthony Lambiase (1975), 'Analysis of high-speed motion pictures of the vocal folds', *Folia Phoniatrica* **27**, 77–87.

Terken, Jacques and Dik Hermes (2000), The perception of prosodic prominence, *in* M. Horne (Ed.), 'Prosody: Theory and Experiment. Studies Presented to Gösta Bruce', Kluwer Academic Publishers, Dordrecht, pp. 89–127.

Thomas, I. B. (1969), 'Perceived pitch of whispered vowels', *Journal of the Acoustical Society of America* **46**(2), 468–470.

Traunmüller, Hartmut and Anders Eriksson (2000), 'Acoustic effects of variation in vocal effort by men, women, and children', *Journal of the Acoustical Society of America* **107**(6), 3438–3451.

Turk, Alice, Satsuki Nakai, and Mariko Sugahara (2006), Acoustic segment durations in prosodic research: A practical guide, *in* Stefan Sudhoff, Denisa Lenertová, Roland Meyer, Sandra Pappert, Petra Augursky, Ina Mleinek, Nicole Richter, and Johannes Schließer (Eds.), 'Methods in Empirical Prosody Research', number 3 *in* 'Language, Context, and Cognition', Mouton de Gruyter, Berlin, pp. 1–26.

Whalen, D.H. and Andrea G. Levitt (1995), 'The universality of intrinsic f_0 of vowels', *Journal of Phonetics* **23**(3), 349–366.

Wouters, Johan and Micael W Macon (2002), 'Effects of prosodic factors on spectral dynamics. I. analysis', *Journal of the Acoustical Society of America* **111**(1), 417–427.

Zemlin, Willard R. (1988), *Speech and Hearing Science: Anatomy and Physiology*, 3rd edn, Prentice-Hall, New Jersey, USA.