University of California
Los Angeles

# The Voice Source in Speech Production: Data, Analysis and Models

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

by

**Yen-Liang Shue**

2010

The dissertation of Yen-Liang Shue is approved.

Mihaela var der Schaar

Kung Yao

Patricia Keating

Abeer Alwan, Committee Chair

University of California, Los Angeles

2010

*To Ma and Ba,*

*the foundations of my existence.*

# Table of Contents

**1 Introduction** . . . . . **1**

1.1 Overview and motivation . . . . . 1

1.2 The linear speech production model . . . . . 6

1.2.1 Voice source models . . . . . 7

1.2.2 The vocal tract . . . . . 10

1.3 Estimation of the voice source signal . . . . . 13

1.4 Voice quality . . . . . 15

1.4.1 Measures related to voice quality . . . . . 16

1.4.2 Prosody . . . . . 21

1.5 Dissertation outline . . . . . 22

**2 A New Voice Source Model based on High-Speed Imaging of the Vocal Folds** . . . . . **23**

2.1 Existing voice source models . . . . . 23

2.2 High-speed imaging of the vocal folds with synchronous audio recordings . . . . . 26

2.2.1 Subjects and voice samples . . . . . 26

2.2.2 Image segmentation . . . . . 27

2.2.3 Glottal area estimation . . . . . 29

2.3 A new voice source model . . . . . 31

2.3.1 Properties of the new source model . . . . . 35

2.3.2 Incomplete glottal closures and the DC-offset . . . . . 35
2.3.3 Glottal area or glottal flow? . . . . . 36
2.4 Evaluation of the new source model . . . . . 37
2.5 Summary and discussion . . . . . 40

**3 A Codebook Search Technique for Estimating the Voice Source 41**
3.1 Voice source estimation . . . . . 41
3.2 Data . . . . . 44
3.3 Method . . . . . 44
3.4 Results . . . . . 52
3.5 Summary . . . . . 58

**4 Acoustic Correlates of Voice Quality . . . . . 60**
4.1 Acoustic measures related to voice quality and to the voice source 60
4.2 `VoiceSauce` - a program for voice analysis . . . . . 63
4.2.1 $F_0$ and formant calculations . . . . . 63
4.2.2 Harmonic magnitudes and spectral amplitude calculations and corrections . . . . . 64
4.2.3 Energy calculation . . . . . 65
4.2.4 *CPP* and *HNR* calculation . . . . . 66
4.3 Application I: Voice quality analysis with respect to acoustic measures . . . . . 66
4.3.1 Data . . . . . 67
4.3.2 Methods . . . . . 67

4.3.3 Results . . . 69
4.3.4 Summary . . . 79
4.4 Application II: Automatic gender classification . . . 80
4.4.1 Speech data . . . 82
4.4.2 Methods . . . 82
4.4.3 Results and discussion . . . 84
4.4.4 Summary . . . 89
4.5 Application III: Prosody analysis . . . 91
4.5.1 Speech corpus . . . 92
4.5.2 Voice quality related measures . . . 94
4.5.3 Contour Fitting and Analysis . . . 94
4.5.4 Results . . . 96
4.5.5 Summary . . . 101
4.6 Summary and discussion . . . 102

**5 Acoustic Correlates of High and Low Nuclear Pitch Accents in American English . . . 104**

5.1 Prosody and pitch accents . . . 104
5.2 Corpus and analysis methods . . . 109
5.2.1 Corpus . . . 109
5.2.2 Analysis methods . . . 111
5.3 Results . . . 114
5.3.1 $F_0$ . . . 116

5.3.2 Energy . . . . . 124
5.3.3 Duration - effects of pitch accent on phrase final lengthening 127
5.4 Discussion . . . . . 132
5.4.1 Overall results . . . . . 135
5.4.2 Individual speaker analyses . . . . . 138
5.4.3 Theories of tonal crowding . . . . . 143
5.5 Conclusion . . . . . 147

**6 Summary and Future Work . . . . . 149**
6.1 Summary . . . . . 149
6.1.1 Source modeling and estimation . . . . . 150
6.1.2 Correlates of voice quality . . . . . 151
6.1.3 Correlates of pitch accents . . . . . 152
6.2 Unsolved issues and outlook . . . . . 153

**A Averaged Glottal Area Waveforms . . . . . 155**

**B Glottal Area Model Fitting Performance of the Proposed New Source Model . . . . . 161**

**C Voice Source Estimation Results for each Subject . . . . . 171**

**References . . . . . 181**

# List of Figures


1.1 Simplified human speech production consisting of the lungs, voice source and vocal tract (from [Fla65]). . . . . 2

1.2 The vocal folds in a (a) closed position and (b) open position. . . 4

1.3 The linear source-filter model of speech production [Fan70]. The top panels show the model in the time domain, and the bottom panels show the model in the frequency domain. . . . . 6

1.4 The linear source-filter model of speech production with the differentiated voice source. . . . . 7

1.5 Example of the Rosenberg model where $T_P/(T_P + T_N) = 0.8$. . . . 9

1.6 Example of the LF model showing the five main parameters: $t_a$, $t_c$, $t_e$, $t_p$, and $E_e$. . . . . 10

1.7 Examples of formant frequencies for /iy/ (solid line) and /æ/ (dotted line). . . . . 12

1.8 Part of a spectrum for an /æ/ vowel showing the LP envelope incorrectly detecting a formant frequency at around 200 Hz. . . . 14

1.9 Magnitude spectrum of the vowel /æ/ showing the measures: pitch frequency ($F_0$), harmonic amplitudes ($H_1$, $H_2$ and $H_4$), formant frequencies ($F_1$, $F_2$ and $F_3$), and the spectral magnitudes at the formant frequencies ($A_1$, $A_2$ and $A_3$). . . . . 17

2.1 The LF model. Top panel illustrates the glottal flow derivative: instant of maximum airflow ($t_p$), instant of maximum airflow derivative ($t_e$), effective duration of return phase ($t_a$), beginning of closed phase ($t_c$), fundamental period $T_0$, and amplitude of maximum excitation of glottal flow derivative ($E_e$). Bottom panel illustrates the glottal flow model. . . . . 25

2.2 Glottal area estimation procedure. Images shown are from the low $F_0$, breathy phonation of subject FM1. . . . . 28

2.3 Glottal area waveform averaging and normalization. Waveforms are from the low $F_0$, breathy phonation of subject FM1. . . . . 30

2.4 The averaged glottal waveforms for the nine phonation combinations for subject FM1. $F_0$ (low, normal and high) was varied quasi-orthogonally with voice quality (pressed, normal and breathy). Note that the three voice quality types differ little in the opening, with most of the difference seen in the closing and in the minimum values. . . . . 32

2.5 Example of the proposed source model with $OQ = 0.7$, $\alpha = 0.6$, $S_{op} = 0.5$ and $S_{cp} = 0.7$. . . . . 34

2.6 Model fitting performance for the phonations from subject FM1 (left panel: low $F_0$, pressed) and subject M2 (right panel: low $F_0$, normal). . . . . 39

2.7 Model fitting performance for the low $F_0$, normal phonation from subject FM3. . . . . 39

3.1 A typical inverse-filtering process. The residual signal is then used to map onto a source model. . . . . 43

3.2 The main source estimation procedure. . . . . 45

3.3 Block diagram showing the method for generating the codebook. Any voice source model can be used to generate the codebook. . . 47

3.4 Block diagram showing the two iterations of the source estimation method; solid and dashed lines represent the first and second iteration, respectively. The codebook sizes are based on the proposed new source model described in Chapter 2. . . . . 51

3.5 MSEs averaged across all phonations for each gender in terms of the voice quality (pressed, normal and breathy) and type of formant constraint (Snack, manual and constant). . . . . 55

3.6 MSEs averaged across all phonations for each gender in terms of the $F_0$ type (low, normal and high) and type of formant constraint (Snack, manual and constant). . . . . 55

3.7 Phonation with the lowest source estimation error (MSE = 0.0018). The measured source waveform was taken from the high $F_0$, pressed phonation of subject FM1. The estimated source waveform (dashed) was from the manual-based formant constraints method. . . . . 56

3.8 Phonation with the highest source estimation error. The measured source waveform was taken from the high $F_0$, breathy phonation of subject FM3 with the DC-offset removed. The dashed line shows the estimated waveform using Snack-based formant constraints (MSE = 0.2995) and the dotted line shows the estimated waveform using constant-based formant constraints (MSE = 0.0116). . 57

4.1 Mean $OQ$ values for each speaker averaged over the pressed, normal and breathy phonations. . . . . 71

4.2 Examples of voice source shapes for the mean $OQ$ and $\alpha$ values listed in Table 4.3; $S_{op}$ and $S_{cp}$ were both set to a value of 0.5. . . 72

4.3 Examples of voice source shapes for the mean $OQ$, $\alpha$ and $S_{op}$ values listed in Table 4.4; $S_{cp}$ was set to a value of 0.5. . . . . 74

4.4 Gender classification accuracy for each age group using just $F_0$, just FB, and $F_0$ plus FB (M0). . . . . 85

4.5 Gender classification accuracy for each age group using the measures sets M1, M2 and M3. M0 represents the baseline performance results. The corresponding values are listed in Table 4.9. . . . . 87

4.6 Average stylized $F_0$ contours "Dagada" (males). . . . . 97

4.7 Average stylized $F_0$ contours for "doodads" (females). . . . . 98

4.8 Average stylized $H_1^* - H_2^*$ contours for "Dagada" (males). . . . . 99

4.9 Stylized $H_1^* - A_3^*$ contours for "Dagada" for a male talker showing syllable boundaries for an instance of each prosodic case. . . . . 100

5.1 Example of polynomial fitting for the target word *dagada* with a high ($H^*$) pitch accent. The top panel shows the waveform, the bottom panel shows the raw and stylized $F_0$ contours. The dotted vertical lines mark the position of the manual segmentation. . . . 112

5.2 Example of polynomial fitting for the target word *dada* with a low ($L^*$) pitch accent. The top panel shows the waveform, the bottom panel shows the raw and stylized $F_0$ contours. The dotted vertical line marks the position of the manual segmentation. . . . . 113

5.3 Scatter plot for the target word *dagada* showing relative $F_0$ peaks for $H^*$ and their relative positions in the accented target vowel for a male speaker in three different contexts: 1. no-bnd-early/no-bnd-early-daily (triangles); 2. bnd (crosses); 3. no-bnd-late-daily (circles). . . . . 119

5.4 Scatter plot for the target word *dada* showing relative $F_0$ peaks for $H^*$ and their relative positions in the accented target vowel for a female speaker in the three different contexts. . . . . 120

5.5 Scatter plot of relative mean vowel energy of the $H^*$ accented vowel for the target word *dagada* for all male speakers in three different contexts: 1. *no-bnd-early/no-bnd-early-daily* (triangles); 2. *bnd* (crosses); 3. *no-bnd-late-daily* (circles). . . . . 125

5.6 Scatterplot of relative mean vowel energy of the $H^*$ accented vowel for the target word dada for all female speakers in three different contexts: 1. *no-bnd-early/no-bnd-early-daily* (triangles); 2. bnd (crosses); 3. *no-bnd-late-daily* (circles). . . . . 126

5.7 Average main-stressed syllable duration with no (Non), $L^*$, and $H^*$ pitch accents for male and female speakers for the target word *dagada* in the early, late and boundary positions. . . . . 129

5.8 Average final vowel durations and error bars of the unaccented words *dagada* and *dada* for male and female speakers in the *late*, *early* and *boundary* positions. The increased duration for the *boundary* case confirms the effects of phrase final lengthening. . . 130

5.9 Average main-stressed vowel durations and error bars of the unaccented words *dagada* and *dada* for male and female speakers in the late, early and boundary positions. The increased duration for the boundary case confirms the unaccented main-stressed syllable lengthening at the boundary condition. . . . . 131

5.10 Average final syllable durations for male speakers for phrase final target word *dagada* with no preceding accents, with a preceding $H^*$ accent, and with a preceding $L^*$ accent. . . . . 133

5.11 Average final syllable durations for female speakers for phrase final target word *dada* with no preceding accents, with a preceding $H^*$ accent, and with a preceding $L^*$ accent. . . . . 134

5.12 Scatter plot for the target word *dagada* showing relative $F_0$ peaks for $H^*$ and their relative positions in the accented target vowel for the male speaker M1 in three different contexts: 1. *no-bnd-early/no-bnd-early-daily* (triangles); 2. *bnd* (crosses); 3. *no-bnd-late-daily* (circles). . . . . 140

5.13 Scatter plot showing the times for the raw peak position in milliseconds from the V onset and the normalized peak heights for the three cases; *no-bnd-early(-daily)* (triangles), *bnd* (crosses) and *no-bnd-late-daily* (circles). The left panel shows the times and heights for a typical male speaker for the target word *dagada* and the right panel shows the results for a typical female speaker for the target word *dada*. . . . . 146

A.1 The averaged glottal waveforms for the nine phonation combinations for subject FM2. $F_0$ (low, normal and high) was varied quasi-orthogonally with voice quality (pressed, normal and breathy). Data for the pressed phonation with normal $F_0$ was not available for this subject. . . . . 156

A.2 The averaged glottal waveforms for the nine phonation combinations for subject FM3. $F_0$ (low, normal and high) was varied quasi-orthogonally with voice quality (pressed, normal and breathy). . . 157

A.3 The averaged glottal waveforms for the nine phonation combinations for subject M1. $F_0$ (low, normal and high) was varied quasi-orthogonally with voice quality (pressed, normal and breathy). Data for the low $F_0$ phonations was not available for this subject. 158

A.4 The averaged glottal waveforms for the nine phonation combinations for subject M2. $F_0$ (low, normal and high) was varied quasi-orthogonally with voice quality (pressed, normal and breathy). . . 159

A.5 The averaged glottal waveforms for the nine phonation combinations for subject M3. $F_0$ (low, normal and high) was varied quasi-orthogonally with voice quality (pressed, normal and breathy). . . 160

B.1 Model fitting performance of the proposed new source model for subject FM1. . . . . 162

B.2 Model fitting performance of the proposed new source model for subject FM2. . . . . 163

B.3 Model fitting performance of the proposed new source model for subject FM3. . . . . 164

B.4 Model fitting performance of the proposed new source model for subject M1. . . . . 165

B.5 Model fitting performance of the proposed new source model for subject M2. . . . . 166

B.6 Model fitting performance of the proposed new source model for subject M3. . . . . 167

C.1 Plot of the measured (solid line) and estimated glottal area waveforms for subject FM1. The estimated waveforms are from the Snack-based (dotted line) and manual-based (dashed line) formant constraints. . . . . 175

C.2 Plot of the measured (solid line) and estimated glottal area waveforms for subject FM2. The estimated waveforms are from the Snack-based (dotted line) and manual-based (dashed line) formant constraints. . . . . 176

C.3 Plot of the measured (solid line) and estimated glottal area waveforms for subject FM3. The estimated waveforms are from the Snack-based (dotted line) and manual-based (dashed line) formant constraints. . . . . 177

C.4 Plot of the measured (solid line) and estimated glottal area waveforms for subject M1. The estimated waveforms are from the Snack-based (dotted line) and manual-based (dashed line) formant constraints. . . . . 178

C.5 Plot of the measured (solid line) and estimated glottal area waveforms for subject M2. The estimated waveforms are from the Snack-based (dotted line) and manual-based (dashed line) formant constraints. . . . . 179

C.6 Plot of the measured (solid line) and estimated glottal area waveforms for subject M3. The estimated waveforms are from the Snack-based (dotted line) and manual-based (dashed line) formant constraints. . . . . 180

# List of Tables

1.1 Some common voice quality related measures. . . . . 18

2.1 Mean $F_0$ values for the low, normal and high $F_0$ with pressed (p), normal (n) and breathy (b) voice qualities. '–' denotes data was not available for a particular phonation. . . . . 27

2.2 Model fitting results for each phonation from each speaker. Results are shown for the proposed/LF models. '–' denotes data was not available for a particular phonation. . . . . 38

3.1 Optimization constraints for formant frequencies for each subject. 50

3.2 Results for each subject and the formant constraint method (Snack, manual and constant based). Values are the MSEs, averaged over all of a subject's phonations, for the LF/proposed new source models. 53

3.3 Correlation coefficients ($r$) for the model-fitted source parameters and the estimated source parameters from the Snack-, manual-, and constant-based formant frequency constraints. The significance levels are in parenthesis, where '–' denotes a particular correlation was not statistically significant. . . . . 54

4.1 List of acoustic measures though to be related to the voice source and/or voice quality . . . . . 62

4.2 Occurrences of glottal gaps in terms of speaker, $F_0$ type (low, normal and high) and voice quality (pressed, normal and breathy). '–' denotes an entry where no speaker produced a glottal gap. . . 68

4.3 Voice source model parameters and acoustic measures which were affected by voice quality in a statistically significant way. Values shown are the $F$ value (ratio of the model mean square to the error mean square), $\eta^2$ (measure of the effect size), and the parameter/measure means and standard deviations (in parentheses) for the three voice qualities. . . . . 70

4.4 Voice source model parameters and acoustic measures which were statistically significant to the effects of the glottal gap. Values shown are the $F$ value, $\eta^2$, and the parameter/measure means and standard deviations (in parentheses) for the phonations with glottal gaps and without glottal gaps. . . . . 75

4.5 Correlations between voice source model parameters and acoustic measures. Values are the correlation coefficients ($r$); correlations with $r > 0.4$ are in bold and were all statistically significant. Measures $H_2-H_4$ and *Energy* did not show any meaningful correlations with any voice source parameters. . . . . 78

4.6 Distribution of gender and utterances for each age group. . . . . 83

4.7 Distribution of utterances used in perception experiments. . . . . 84

4.8 Measure sets (M0–M3) used in the gender classification tests. M0, in bold, is used as the baseline measure set. . . . . 86

4.9 Gender classification accuracy for the different measurement sets (M0-M3) and age groups. MFCC feature classification results are shown for comparison. . . . . 88

4.10 Gender classification accuracy for age group 12-13, distinguishing between males and females. . . . . 88

4.11 SVM gender classification accuracy, in percent, using measure set M2 compared with perception results from this paper (PER1) and from Perry et al. [POA01](PER2). Dashes indicate unavailable values. The perception experiments used the target words. . . . . 90

5.1 Position of the $F_0$ peak/trough as a percentage of the speaker's vowel duration. The results shown are averaged for the male and female speakers for the target words *dagada* and *dada*; standard deviation values are shown in parentheses. The statistical significance (s.s.) column shows the ANOVA results for *no-bnd* vs. *bnd*. For significant results, the $F$ (ratio of the model mean square to the error mean square) and $\eta^2$ (measure of the effect size) values are given. . . . . 116

5.2 Height of the $F_0$ excursion as a percentage of the speaker's mean $F_0$. Average results are shown for the *no-bnd* vs. *bnd* conditions for the male and female speakers for the target words *dagada* and *dada*; standard deviation values are shown in parentheses. . . . . 117

5.3 Position of the $F_0$ peak/trough as a percentage of the speaker's target vowel duration. Results shown are average values for the male and female speakers for target words *dagada* and *dada* in the *no-bnd-early(-daily)* vs. *bnd* condition; standard deviation values are shown in parentheses. . . . . 121

5.4 Height of the $F_0$ excursion as a percentage of the speaker's mean $F_0$. Results shown are for the *no-bnd-early(-daily)* vs. *bnd* conditions; standard deviation values are shown in parentheses. . . . . 122

5.5 Relative position of the $F_0$ peak/trough as a percentage of the speaker's target vowel duration. Results shown are average vales for the male and female speakers for target words *dagada* and *dada* in the *no-bnd-late-daily* vs. *bnd* condition; standard deviation values are shown in parentheses. . . . . 123

5.6 Relative height of the $F_0$ excursion as a percentage of the speaker's mean $F_0$. Results shown are for *no-bnd-late-daily* vs. *bnd*; standard deviation values are shown in parentheses. . . . . 124

5.7 Relative energy mean, standard deviation (std.) in parenthesis, of stressed syllables for the target word *dagada* for male and female speakers. Results are shown for *no-bnd* vs. *bnd*. . . . . 127

5.8 Relative energy mean, standard deviation (std.) in parenthesis, of stressed syllables for the target word dada for male and female speakers. Results are shown for *no-bnd* vs. *bnd*. . . . . 128

5.9 Average duration, standard deviation (in parenthesis) of the final syllable of *dagada*/*dada* with no preceding pitch accent, with a $H^*$ preceding accent, and with a $L^*$ preceding accent. All results were statistically significant. . . . . 135

5.10 Comparison of the speakers M1, M2 and F9's $F_0$ peak position and relative height consistencies with the general trends for the *bnd* case; a 'Yes'/'No' denotes agreement/disagreement while 'N/A' means no enough data was available. . . . . 141

B.1 Voice source parameters from the model fit (see Figure B.1) for subject FM1. "G. gap" denotes the existence/absence of the glottal gap. . . . . 168

B.2 Voice source parameters from the model fit (see Figure B.2) for subject FM2. "G. gap" denotes the existence/absence of the glottal gap. Pressed, normal $F_0$ phonations were not available for this subject. . . . . 168

B.3 Voice source parameters from the model fit (see Figure B.3) for subject FM3. "G. gap" denotes the existence/absence of the glottal gap. . . . . 169

B.4 Voice source parameters from the model fit (see Figure B.4) for subject M1. "G. gap" denotes the existence/absence of the glottal gap. Low $F_0$ phonations were not available for this subject. . . . . 169

B.5 Voice source parameters from the model fit (see Figure B.5) for subject M2. "G. gap" denotes the existence/absence of the glottal gap. . . . . 170

B.6 Voice source parameters from the model fit (see Figure B.6) for subject M3. "G. gap" denotes the existence/absence of the glottal gap. . . . . 170

C.1 MSE values for source estimation using Snack-based formant constraints with the proposed new source model; results listed in terms of voice quality (pressed, normal and breathy) and $F_0$ type (low, normal and high). '–' denotes data was not available for a particular phonation. . . . . 172

C.2 MSE values for source estimation using manual-based formant constraints with the proposed new source model; results listed in terms of voice quality (pressed, normal and breathy) and $F_0$ type (low, normal and high). '–' denotes data was not available for a particular phonation. . . . . 173

C.3 MSE values for source estimation using constant-based formant constraints with the proposed new source model; results listed in terms of voice quality (pressed, normal and breathy) and $F_0$ type (low, normal and high). '–' denotes data was not available for a particular phonation. . . . . 174

# Acknowledgments

I would like to express my deepest gratitude to my advisor, Professor Abeer Alwan for her guidance and support during my time in her lab. Her understanding, patience and wisdom made it easier for me to navigate past the many research obstacles I faced. I would also like to thank my committee members, professors Mihaela van der Schaar, Kung Yao and Patricia Keating for their interest in my work.

Special thanks must be given to professors Patricia Keating and Jody Kreiman, who introduced me to linguistics and statistics. Working with them taught me many things about analysis methods and how to apply them to different data. Their enthusiasm and accessible nature meant that I always had a place to discuss matters of voice quality and the intricacies of the vocal folds.

I was incredibly fortunate and privileged to have been able to collaborate with professors Stefanie Shattuck-Hufnagel, Nanette Veilleux, and Sun-Ah Jun. Their help with designing and recording the prosodically-balanced corpus allowed me to concentrate on the analysis, while their vast knowledge and brilliant insights into the mysteries of prosody often helped me to get unstuck when I was running low on ideas.

One of the things that I realized during my research, was that without good data, no analyses can be performed and no hypotheses tested. To this end, I am greatly indebted to the students from the UCLA Department of Linguistics and also the students from MIT who volunteered to be our subjects. Much appreciation also goes to the glottal bureaucrats at the UCLA Bureau of Glottal Affairs for their invaluable high-speed glottal imaging data.

Finally, I would like to express my gratitude to the two ex-lab members

who were instrumental in helping me start and advance my research: Sankaran "Panchi" Panchapagesan, whose reasoning ability meant that I was never lost in a train of thought, and Markus Iseli, who introduced me to the voice source and was my partner in crime during our many paper-writing escapades.

This research was supported in part by the NSF and by NIH grant DC01797. Parts of this dissertation have appeared in the publications listed under "Publications and Presentations".

# Vita

| | |
|---|---|
| 1980 | Born, Taipei, Taiwan. |
| 12/1998–2/1999 | Development of an interface to an application-specific integrated circuit.<br>Texas Instruments, Taipei, Taiwan. |
| 12/2000-2/2001 | Recipient of the CSE Computer Systems Research Group Summer Scholarship.<br>University of New South Wales, Australia. |
| 2002 | B.E. in Computer Engineering.<br>University of New South Wales, Australia. |
| 2004 | M.S. in Electrical Engineering.<br>University of California, Los Angeles, (UCLA). |
| 2008 | Recipient of the Borgstrom Graduate Scholarship for Speech Research.<br>University of California, Los Angeles, (UCLA). |
| 6-9/2008 | Research and development in the field of speech and audio subsystems.<br>Qualcomm Incorporated, San Diego. |
| 9/2008 | Best Student Paper Award at Interspeech 2008.<br>Brisbane, Australia. |
| 2006–2009 | Research/Teaching Assistant<br>Electrical Engineering Department,<br>University of California, Los Angeles, (UCLA). |

## Publications and Presentations

Y.-L. Shue and A. Alwan, "A new voice source model based on high-speed imaging and its application to voice source estimation," *accepted for IEEE ICASSP*, Dallas, TX, March 2010.

Y.-L. Shue, S. Shattuck-Hufnagel, M. Iseli, S.-A. Jun, N. Veilleux, and A. Alwan, "On the acoustic correlates of high and low nuclear pitch accents in American English", *Speech Communications*, vol. 52, pp. 106–122, 2010.

Y.-L. Shue, P. Keating, and C. Vicenik, "VoiceSauce: A program for voice analysis," *The Journal of the Acoustical Society of America*, San Antonio, TX, vol. 124, no. 4, p. 2221, October 2009.

P. Keating and Y.-L. Shue, "Voice quality variations with fundamental frequency in English and Mandarin," *The Journal of the Acoustical Society of America*, San Antonio, TX, vol. 124, no. 4, p. 2221, October 2009

Y.-L. Shue, J. Kreiman, and A. Alwan, "A novel codebook search technique for estimating the open quotient," *Proceedings of Interspeech*, Brighton, UK, pp. 2895–2898, August 2009.

Y.-L. Shue, S. Shattuck-Hufnagel, M. Iseli, S.-A. Jun, N. Veilleux, and A. Alwan, "Effects of intonational phrase boundaries on pitch-accented syllables in

American English", *Proceedings of Interspeech*, Brisbane, Australia, pp. 873–876, September 2008.

J. Kreiman, B. Gerratt, M. Iseli, J. Neubauer, Y.-L. Shue, and A. Alwan, "The relationship between open quotient and $H1^* - H2^*$," *The Journal of the Acoustical Society of America*, Miami, FL, vol. 124, no. 4, p. 2495, October 2008

J. Kreiman, B. Gerratt, M. Iseli, J. Neubauer, Y.-L. Shue, and A. Alwan, "The relationship between open quotient and $H1^* - H2^*$," *Proceedings of the 6th International Conference on Voice Physiology and Biomechanics*, Tampere, Finland, August 2008.

Y.-L. Shue, M. Iseli, S. Shattuck-Hufnagel, N. Veilleux, S.-A. Jun, and A. Alwan, "Effects of boundary tones on accent-related F0 peak alignment," *The Journal of the Acoustical Society of America*, Paris, France, vol. 123, no. 5, p. 3460, May 2008.

Y.-L. Shue and M. Iseli, "The role of voice source measures on automatic gender classification," *Proceedings of IEEE ICASSP*, Las Vegas, NV, pp. 4493–4496, March 2008.

Y.-L. Shue, M. Iseli, N. Veilleux, and A. Alwan, "Pitch accent versus lexical stress: quantifying acoustic measures related to the voice source," *Proceedings of Interspeech*, Antwerp, Belgium, pp. 2625–2628, August 2007.

M. Iseli, Y.-L. Shue, and A. Alwan, "Age, sex and vowel dependencies of acous-

tical measures related to the voice source," *The Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2283–2295, 2007.

M. Iseli, Y.-L. Shue, M. Epstein, P. Keating, J. Kreiman, and A. Alwan, "Voice source correlates of prosodic features in American English: a pilot study," *Proceedings of Interspeech*, Pittsburgh, PA, pp. 2226–2229, September 2006.

M. Iseli, Y.-L. Shue, and A. Alwan, "Age- and gender-dependent analysis of voice source characteristics," *Proceedings of IEEE ICASSP*, Toulouse, France, vol. 1, pp. 389–392, May 2006.

M. Iseli, Y.-L. Shue, and A. Alwan, "Analysis of vowel and speaker dependencies of source harmonic magnitudes," *The Journal of the Acoustical Society of America*, Vancouver, Canada, vol. 117, no. 4, p. 2619, May 2005.

H. ElGindy and Y.-L. Shue, "On sparse matrix vector multiplication with FPGA-based systems," *Proceedings of the 10th Annual IEEE Symposium on Field-Programmable Custom Computing Machines*, Napa Valley, CA, pp. 273-274, September 2002.

Abstract of the Dissertation

# The Voice Source in Speech Production: Data, Analysis and Models

by

**Yen-Liang Shue**

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2010

Professor Abeer Alwan, Chair

Analysis of the voice source with respect to voice quality is essential to the understanding of the human speech production system, which can lead to better speech modeling for improving a vast range of applications. However, due to the position of the vocal folds, analyzing the source is often hampered by the lack of direct observations with which to calibrate algorithms.

In this dissertation, two approaches to voice source and voice quality analysis were pursued. In the first approach, the source waveform was extracted by analyzing the glottal area waveforms from high-speed imaging of the vocal folds. These direct observations led to the development of a new source model, which is more accurate compared to existing models. A codebook search technique was then proposed to estimate the source signal from the acoustic data. Results were promising for a number of model parameters such as the open quotient and speed of opening. However, error analysis showed that the algorithm required reasonable formant-frequency constraints which may be difficult to obtain automatically in some cases.

In the second approach, voice source related measures were used in three voice

quality applications: voice source analysis, automatic gender classification and prosody analysis. In voice source analysis, acoustic measures were examined in the context of the voice source model parameters obtained from model-fitting the glottal area waveforms. Results showed that correlations could be made between model parameters and the related acoustic measures, such as the asymmetry coefficient and harmonic-to-noise ratio measures. It was also shown that the model parameters and related acoustic measures were affected by the type of voice quality (pressed, normal and breathy). In gender classification, voice source related measures were found to be more helpful in younger (10–14 year old) speakers, where traditional pitch and formant frequency features were less useful. Analysis of prosody showed that, amongst other things, features correlated to pitch accents were not necessarily centered at the target syllable, and depended on the position of other prosodic events.

# CHAPTER 1

# Introduction

## 1.1 Overview and motivation

The human speech production mechanism consists of a system of articulators which, when used harmoniously, allows a speaker to produce a vast range of sounds. The system is physiologically complex, but as shown in Figure 1.1, it can be broadly classified into three main components: the lungs, the voice source and the vocal tract. In simplified terms, the lungs can be thought of as an air pump providing the necessary airflow to stimulate the voice source and vocal tract, allowing the former to dictate how something is being said, i.e. the *voice quality*, and the latter to control what is being said. The term voice quality (at least for English) encompasses a wide range of voice characteristics ranging from whispery to breathy, from lax to tense, from creaky to falsetto, from stressed to non-stressed, and from low-pitched to high-pitched. In order to understand how different voice qualities are produced, it is necessary to delve into the properties of the voice source.

Physiologically, the voice source is created when airflow from the lungs is pushed through the larynx, which is a structure made of cartilage and muscle. The larynx is positioned just above the trachea and its main cartilage is commonly known as the voice box or "Adam's apple". Within the larynx are a pair of muscles which form the vocal folds (also known as the vocal cords). In the speech-

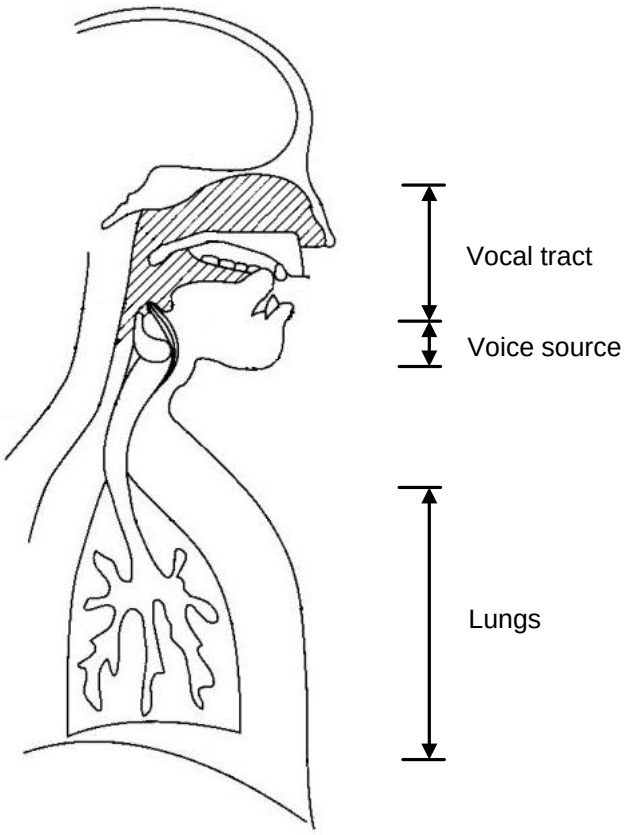
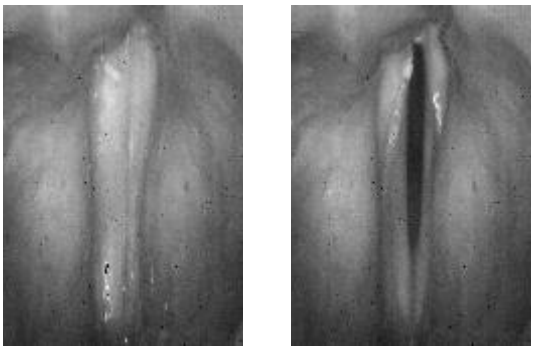



**Figure 1.1:** Simplified human speech production consisting of the lungs, voice source and vocal tract (from [Fla65]).

ready position, the vocal folds are usually closed, as shown in Figure 1.2(a). To produce unvoiced sounds, the vocal folds are held apart, allowing the air unobstructed flow. With a sufficiently high rate of airflow, turbulence is created and a noisy sound is produced. To produce voiced sounds, the muscles which control the closing of the vocal folds (adductor muscles) are used to provide resistance to the air pressure coming from the lungs. Once the pressure has forced the vocal folds to open (Figure 1.2(b)), air rushes through, which decreases the pressure between the folds ("Bernoulli Effect"), causing them to return to the closed position. This cycle is repeated many times during one second and the duration of each cycle is known as the fundamental period ($T_0$). The fundamental frequency ($F_0$) is defined as $F_0 = 1/T_0$ and is commonly referred to as the "pitch frequency" or simply "pitch". Phonation or voicing is a sustained oscillation of the vocal folds, and the rate of this vibration is heard as the pitch of the voice. The rate can be varied, in which case the pitch varies. An animation of the vocal fold vibrations can be seen at `http://www.humnet.ucla.edu/humnet/linguistics/faciliti/demos/vocalfolds/vocalfolds.htm`.

There are considerable physiological differences between the vocal folds of adult males and adult females. In general, adult males have longer and thicker vocal folds which result in a lower $F_0$ value, typically around 100–130 Hz; the average female $F_0$ value is approximately 200–230 Hz.

The pitch frequency is one of the most important parameters of the voice source and it is also the easiest to estimate from speech signals. In English, varying the $F_0$ during speech can be used to signal changes in voice quality; e.g. raising the $F_0$ towards the end of a sentence can cast that sentence as a question or as an expression of surprise. In tonal languages, such as Mandarin, changes in $F_0$ on a word usually result in a totally different meaning.



(a) (b)

**Figure 1.2:** The vocal folds in a (a) closed position and (b) open position.

There are inherent difficulties with analyzing an articulator which cannot be accessed or measured easily. Typically, indirect measurement techniques, such as inverse-filtering (see Section 1.3), are used to infer an estimate of the voice source. Often, these techniques require a good model of the voice source to fit to the inverse-filtered signal, which is somewhat contradictory. Another non-invasive method of obtaining voice source measurements is through the use of electroglottography (EGG) which, in theory, measures the changes in contact between the vocal folds. This is achieved by placing two electrodes on the neck, positioned either side of the larynx, and passing a small high-frequency current through them. The measured resistance waveforms during a voiced phonation reflect the movement of the vocal folds and hence, can be used as a measure of vocal fold contact. However, EGG signals are, at best, another form of indirect measurement, and can be influenced by various factors such as the thickness of the neck and the positioning of the electrodes. In [Rot73], a special mask for measuring the air flow at the mouth and nose was devised. This device was able to measure the absolute air flow, including the DC component. However, a later study [HG92] found that the mask was bandlimited to approximately 1.6 kHz and performed with less accuracy at lower frequencies. Despite the difficulties,

all of the existing voice source models and measurements do show convergence to waveforms which are of a similar shape. However, without empirical data of the "ground truth", it is difficult to ascertain the accuracy of a particular measurement or model.

Instead of directly estimating the voice source to study voice quality, one can also use measures (see Section 1.4.1) which are correlated with some feature of the voice source. These measures can be spectral or temporal and some have been used extensively in studies of voice quality [Fis67, HCE94, Han97, Bla97, HC99, HdD01, CHC05, Esp06, ISA07]. However, there has been less research into how these measures can be affected by factors such as gender and certain prosodic events, such as pitch accents and boundary tones (see Section 1.4.2). This issue is further confounded by cases where multiple prosodic events happen within a short duration, as can often occur in English. While voice source related measures have been widely used, presently, there are very few studies which show how these measures relate physiologically to the vocal-fold movements.

A better understanding of the voice source would improve our knowledge of how voice quality is produced, and how it is affected by speaker and various speech sounds. This knowledge can be applied to many applications such as speech/speaker recognition, speech synthesis, emotion identification, age identification, speech coding and various medical applications.

This dissertation has two main goals. The first seeks to expand our knowledge of the voice source by analyzing direct measurements of the vocal folds. A new source model and source estimation technique is then proposed based on these direct measurements. The second goal is to examine the role of voice source related measures in signaling prosodic events, gender, and the movement of the vocal folds. In the process, a new software application (`VoiceSauce`) was created

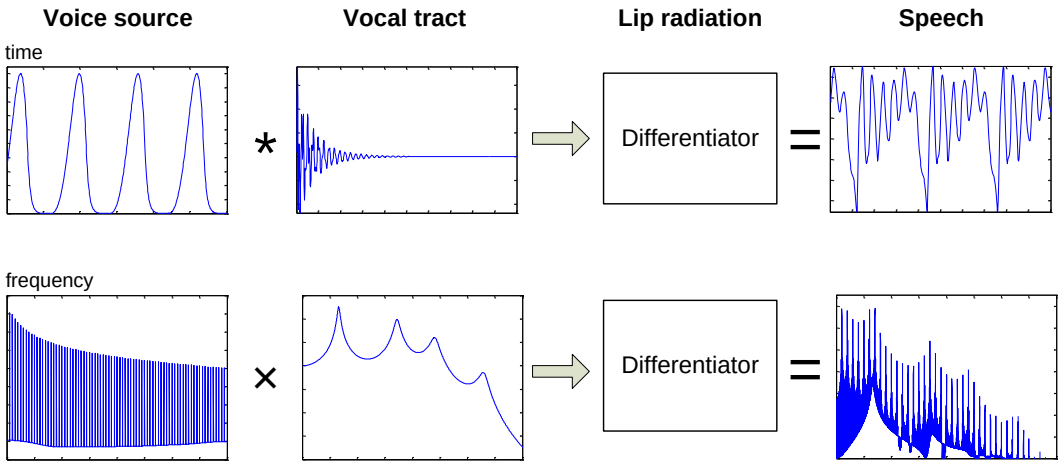



**Figure 1.3:** The linear source-filter model of speech production [Fan70]. The top panels show the model in the time domain, and the bottom panels show the model in the frequency domain.

to simplify the calculation and analysis of these measures.

## 1.2 The linear speech production model

Although speech production is generally a non-linear process, for short time frames, it can be reasonably approximated as a cascade of linear systems involving a source function (voice source), a pole-zero filter (simulates the vocal tract) and a differentiator (simulates lip radiation). This process, shown in Figure 1.3, is known as the linear source-filter model of speech production [Fan70] and is widely used in speech research and applications. Physiologically, non-linear interactions between the vocal tract and the voice source do occur during speech production, but these interactions are not represented by this model. Mathematically, if the speech signal is denoted by $s(t)$, and the source, vocal tract and lip radiation were $u(t)$, $v(t)$ and $r(t)$ respectively, then $s(t) = u(t) * v(t) * r(t)$ and in the spectral domain, $S(\omega) = U(\omega) \cdot V(\omega) \cdot R(\omega)$.

Since a differential operator is often used to simulate lip radiation, it is common practise to move the operator so that it is applied to the source signal, as shown in Figure 1.4. It is important to note that the differentiator is only a simplification and, as noted in [Ste00], this simplification can become inaccurate at very low frequencies.

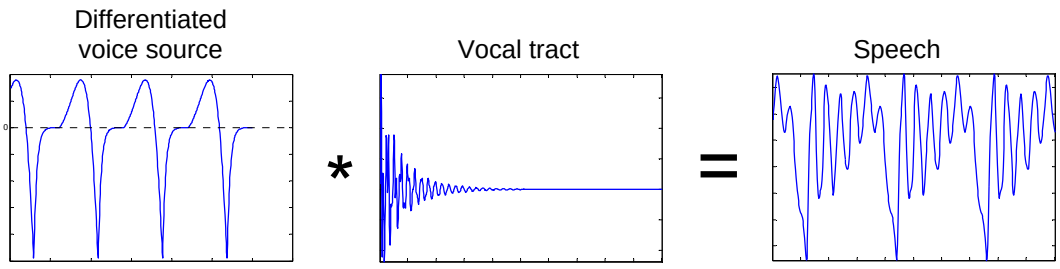



**Figure 1.4:** The linear source-filter model of speech production with the differentiated voice source.

### 1.2.1 Voice source models

Voice source models can be broadly categorized into two main types: interactive models which formally describe interaction between the source and the vocal tract, and non-interactive models, which assume linear source/tract interactions. Interactive models are generally more complex and involve calculating the glottal flow signal in relation to the different coupling effects of the voice production system. However, since these effects are not well understood, non-interactive models have provided a popular alternative.

Many non-interactive models with varying complexities have been proposed, such as the Rosenberg [Ros71], Hedelin [Hed84], Fant [Fan79], Ananthapadmanabha [Ana84], Liljencrants-Fant (LF) [FLL85], and Fujisaki-Ljungqvist [FL86]

models. The motivations for such a wide range of models are mainly due to the different types of data and observations on which the models are built. These observations have come from air-flow masks, EGG signals, mechanical systems, and inverse-filtering of speech signals based on the linear speech production model [Fan70]. The simplest of the aforementioned source models, such as the Rosenberg, Hedelin, and Fant models, are based on sinusoidal functions, while the Ananthapadmanabha, LF, and Fujisaki-Ljungqvist models use more complicated combinations of sinusoids, exponentials and polynomials. A summary of the differences of these models is presented here, but more detailed analyses can be found in [CC95] and [FL86].

The Rosenberg models are probably the simplest and easiest to generate. In [Ros71], many source models were synthesized and perceptually tested for naturalness. It was found that the two models which were judged to be the most natural, had pulses which were similar to waveforms that had been derived from the inverse-filtering of speech signals. The two models were similar in shape, but one was generated from polynomials while the other was generated using trigonometric functions. The Rosenberg trigonometric source model is a glottal flow model with three parameters and, is defined as:

$$u_g(t) = \begin{cases} \frac{\alpha}{2}\left(1 - \cos\left(\frac{\pi t}{T_P}\right)\right), & 0 \leq t \leq T_P \\ \alpha \cos\left(\frac{\pi(t-T_P)}{2T_N}\right), & T_P < t \leq T_P + T_N \\ 0, & \text{during glottal closure} \end{cases} \tag{1.1}$$

where $\alpha$ is the maximum amplitude of the glottal pulse, $T_P$ is the time from the glottal pulse onset to the maximum amplitude and $T_N$ is the time from the maximum amplitude to the glottal pulse offset. This model, an example of which is shown in Figure 1.5, has two separate functions for the opening and closing phases. The Hedelin model was developed based on the Rosenberg trigonometric

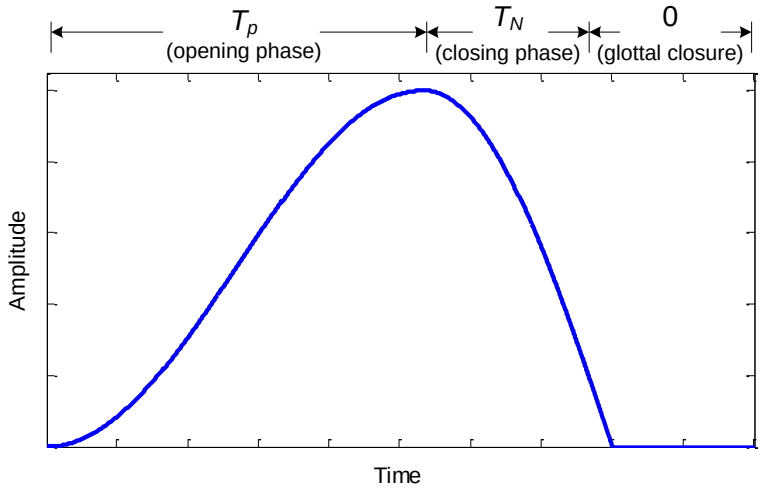



**Figure 1.5:** Example of the Rosenberg model where $T_P/(T_P + T_N) = 0.8$.

model and was used primarily for analysis and synthesis in a LPC-based vocoder. The main difference was the addition of a low frequency drift component which, unlike the Rosenberg model, produced non-zero values during the glottal closure. The Fant model proposed using functions, similar to the Rosenberg trigonometric model, but with the ability of controlling the flow derivative discontinuity. The Ananthapadmanabha model further refined the Fant model using data derived from inverse-filtering. Unlike the previous models, the Ananthapadmanabha model was specified as a derivative flow waveform, which incorporated the lip radiation effects into the source model. Similarly, the LF (shown in Figure 1.6) and Fujisaki-Ljungqvist models were also specified as derivative flow models. With its six parameters, the Fujisaki-Ljungqvist model provided greater depth in its modeling ability. The use of polynomials instead of trigonometric functions further allowed the number of parameters to be varied according to the required level of detail. However, the large number of parameters also made it difficult to use in analysis and synthesis, in that not all parameter combinations are able to generate a valid source waveform.

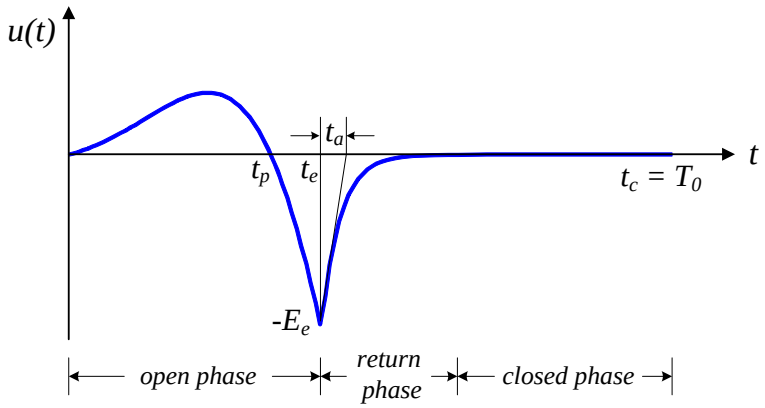



**Figure 1.6:** Example of the LF model showing the five main parameters: $t_a$, $t_c$, $t_e$, $t_p$, and $E_e$.

Although there are many different voice source models, their basic shapes are similar and appear approximately like what is shown in Figure 1.5. The major differences are in the handling of the pulse onset, offset and in the tilt of the pulse.

### 1.2.2 The vocal tract

The vocal tract functions like an adjustable tube, consisting of many articulators such as the tongue, palate, nasal cavities, teeth, and lips. Different sounds are produced when the shape of this tube is altered by the movement of these articulators. For example, the vowel /iy/ as in "bead" has a much smaller mouth opening than the vowel /æ/ as in "bat". The vocal tract can also be used to produce turbulent noise for unvoiced sounds by forcing the air to flow through a narrow channel. The fricative /s/ as in "set" is one such example where the mouth opening is almost closed, with the tongue positioned close to the palate.

In the frequency domain, the vocal tract has the effect of shaping the source spectrum. For voiced speech, this shape consists of resonances (*formants*) and

anti-resonances. The formants dictate how much energy from the voice source is transferred to the lips and each formant is described by its *formant frequency*, that is, the frequency of resonance and its *formant bandwidth*, the resonance bandwidth. Anti-resonances represent energy loss and are typically produced mainly during the production of consonants. Different sounds have different formant frequencies. For example, in [PB52] it was found that the average values of the first three formant frequencies for adult male talkers for the vowel /iy/ were at 270 Hz, 2290 Hz, and 3010 Hz, whereas the average formant frequencies for the vowel /æ/ were at 660 Hz, 1720 Hz and 2410 Hz. This is shown in Figure 1.7 using ideal vocal tract shapes; $F_1$, $F_2$ and $F_3$ denote the first, second and third formant frequencies, respectively.

The vocal tract is usually modeled as a passive acoustic filter containing poles and zeros. This can be expressed as:

$$V(z) = G\frac{B(z)}{A(z)}$$

where $G$ is the gain factor, $A(z)$, the poles of $V(z)$, represent the formants, and $B(z)$, the zeros of $V(z)$ represent the anti-resonances of the vocal tract. For vowels, which carry most of the energy in speech signals, the filter mainly consists of poles, where each complex-conjugate pole-pair represents a formant frequency and its corresponding bandwidth. Mathematically, the transfer function of an all-pole filter can be expressed in the $\mathcal{Z}$-domain as:

$$V(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{1.2}$$

where $p$ is the order of poles and $a_k$'s are the coefficients. A set of linear equations are solved by using linear prediction (LP) techniques [Mak75a] to obtain the $a_k$ values.

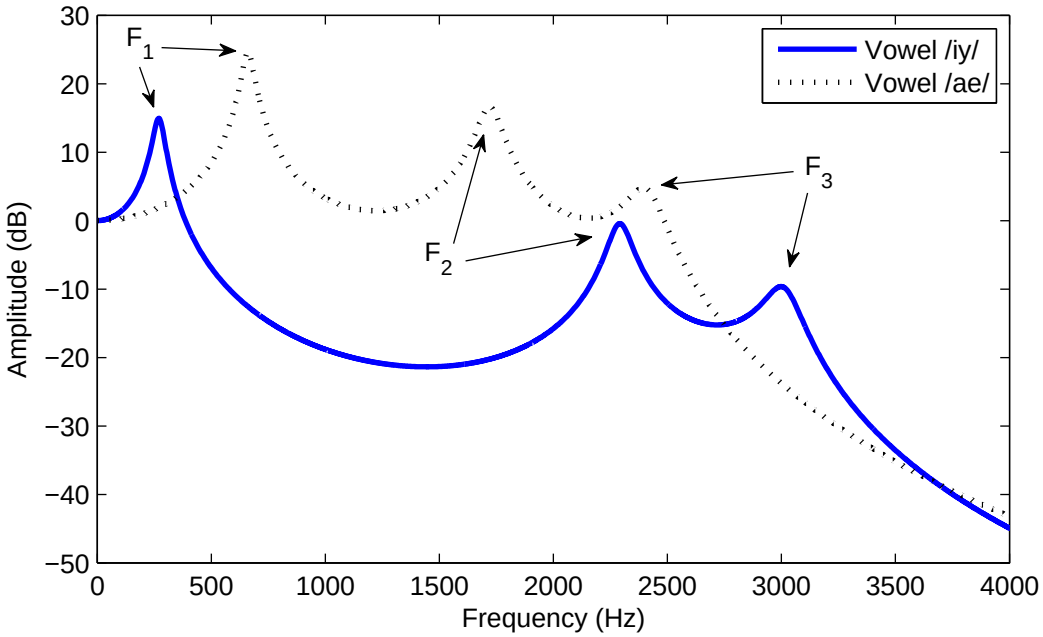



**Figure 1.7:** Examples of formant frequencies for /iy/ (solid line) and /æ/ (dotted line).

## 1.3 Estimation of the voice source signal

From the linear speech production model, speech, $s(t)$ is the result of an interaction between between the vocal tract, $v(t)$, and the (differentiated) voice source, $u(t)$: $s(t) = v(t) * u(t)$. Hence, in the $\mathcal{Z}$-domain, the voice source could theoretically be estimated if the vocal tract parameters from Eq. 1.2 were known:

$$U(z) = \frac{S(z)}{V(z)} \tag{1.3}$$

In practice, voice source estimation is significantly more complex due to the interactions between the vocal tract and the source.

Eq. 1.3 forms the basis for *inverse-filtering*, where the inverse of the vocal tract transfer function (VTTF) is used to filter the speech signal in order to obtain an estimate of the voice source. This requires a good estimate of the VTTF which is itself a non-trivial task. Typical LP-based methods can introduce inaccuracies for high-pitched voices. As shown in [Mak75b], these inaccuracies are mainly due to the error function used to perform the LP:

$$E_{LP} = \frac{1}{N} \sum_{n=1}^{N} \frac{P(\omega_n)}{\hat{P}(\omega_n)} \tag{1.4}$$

where $N$ is the number of frequency bins, $P(\omega)$ is the power spectrum of the actual signal at frequency $\omega_n$ and $\hat{P}(\omega_n)$ is the estimated power spectrum at $\omega_n$. This error criterion is not unbiased, and favors $\hat{P}(\omega_n)$ values which are greater than $P(\omega_n)$; in terms of speech spectrum matching, this means the valleys in the actual spectrum will tend to be overestimated while the poles in the estimated spectrum will be centered around the peaks of the original speech spectrum. In [EM91], a new error criterion was used based on the Itakura-Saito distortion measure [IS68]. While it was shown in that study that the new error criterion produced more accurate results than the standard LP criterion (Eq. 1.4), there

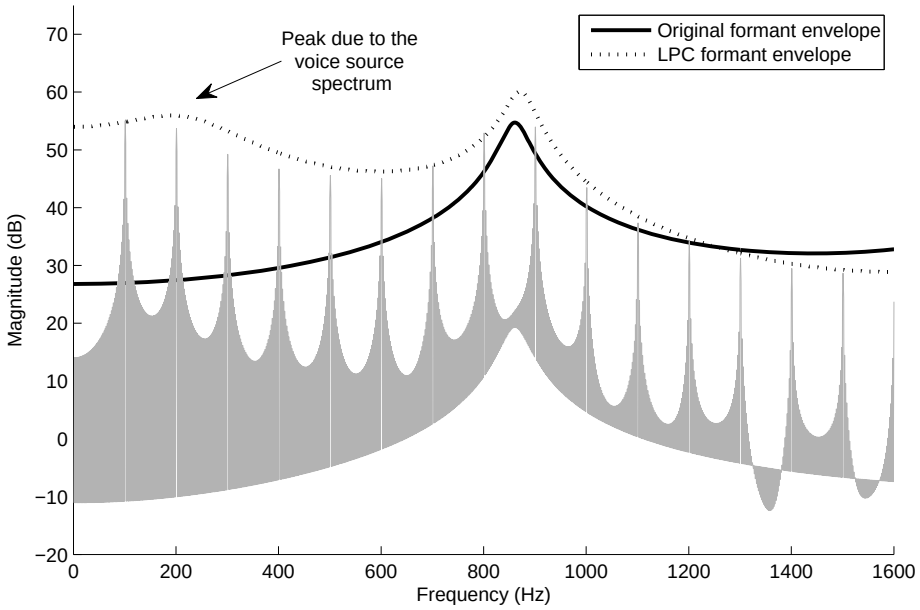



**Figure 1.8:** Part of a spectrum for an /æ/ vowel showing the LP envelope incorrectly detecting a formant frequency at around 200 Hz.

remains a fundamental flaw with estimating the VTTF before estimating the voice source. Figure 1.8 shows the spectrum for a synthesized /æ/ vowel with the LP spectrum envelope and the original VTTF envelope. Because the position of the first formant is so high, at around 850 Hz, the voice source spectrum is able to dominate the lower frequencies, resulting in a "false" peak at around 200 Hz. Currently, there are two main methods of addressing this issue: (1) VTTF estimation can be performed when the interactions between the voice source and the vocal tract are at a minimum, or (2) by using joint source-tract estimation.

In normal phonation, the voice source is usually off during a certain proportion of time, commonly referred to as the *glottal closure regions*. During these regions, little interaction between the voice source and the VTTF occurs and, in

theory, a better estimate of the VTTF can be obtained. Naturally, this method requires first determining the position of the glottal closure regions, which is not an entirely solved problem. In [IC04], sliding windows with variable lengths were used to estimate the VTTF by successively narrowing down the glottal closure regions. Inverse-filtered speech signals were compared with EGG signals from 2 male and 2 female subjects, and the results were quite agreeable. However, in high-pitched voices and certain types of phonations, such as breathy voices, there may be little or no glottal closure, which would affect the accuracy of this method to estimate the voice source.

Joint source-tract estimation methods focus on estimating the VTTF and voice source together. In [FMS01], the LF source model parameters were estimated iteratively using multi-dimensional optimization techniques that were initialized based on the results of an exhaustive parameter search. During the exhaustive search, a source parameter set was tested by removing its spectrum from the speech spectrum, estimating the VTTF and comparing the output of inverse-filtering with the source spectrum from the parameter set. Multi-dimension optimizers were then employed to further refine the parameters. In [PY08], [JI05] and [PB09], a global optimization scheme was used to estimate the parameters of the source and VTTF simultaneously. In these studies, the results of the source estimation were compared with EGG signals and found to be in reasonable agreement.

## 1.4 Voice quality

Voice quality refers to the component of speech which characterizes, either temporarily or permanently, a speaker's voice or speaking style. Many of these characteristics come from the movement patterns of the vocal folds and have

traditionally (e.g. [CG97]) been described as having three main linguistic modes: modal, breathy and pressed. Modal voices are the most common and occur when the vibrations of the vocal folds are periodic with full closing of the glottis. In this mode, very little friction noise is produced when air flows through the glottis and the resulting spectrum has little high frequency components. Breathy voices are produced when a large volume of air is expended during phonation. An effect of this is that sometimes the glottis may not be fully closed during the vocal fold vibrations. Pressed phonations are usually characterized by tense vocal folds and smaller glottal openings. Other modes are possible such as tense, lax, slack, stiff, and creaky.

### 1.4.1 Measures related to voice quality

Instead of directly estimating the voice source signal to study voice quality, one can also look at measures which are related to the voice source or voice quality. These measures are usually taken from the speech spectrum, an example of which is shown in Figure 1.9 for the vowel /æ/ labeled with some of the commonly used measures: the pitch frequency ($F_0$), the amplitudes of the harmonics ($H_1$, $H_2$ and $H_4$), and the spectral amplitudes at the formant frequencies ($A_1$, $A_2$ and $A_3$). Combinations of these measures, such as $H_1 - H_2$, are usually used to normalize the effects of the signal power. Corrections to remove the influences of the VTTF can also be employed so that the measures are more related to the voice source signal; corrected measures are usually denoted by an asterisk ('*') as in $H_1^* - H_2^*$. Table 1.1 lists some of the commonly used voice source related measures.

$F_0$ is the most widely used parameter in the study of voice quality. By definition, pitch accents are rapid excursions of the voice pitch from the "normal" $F_0$ range of a speaker. These excursions can be of a high (above normal) or low (be-

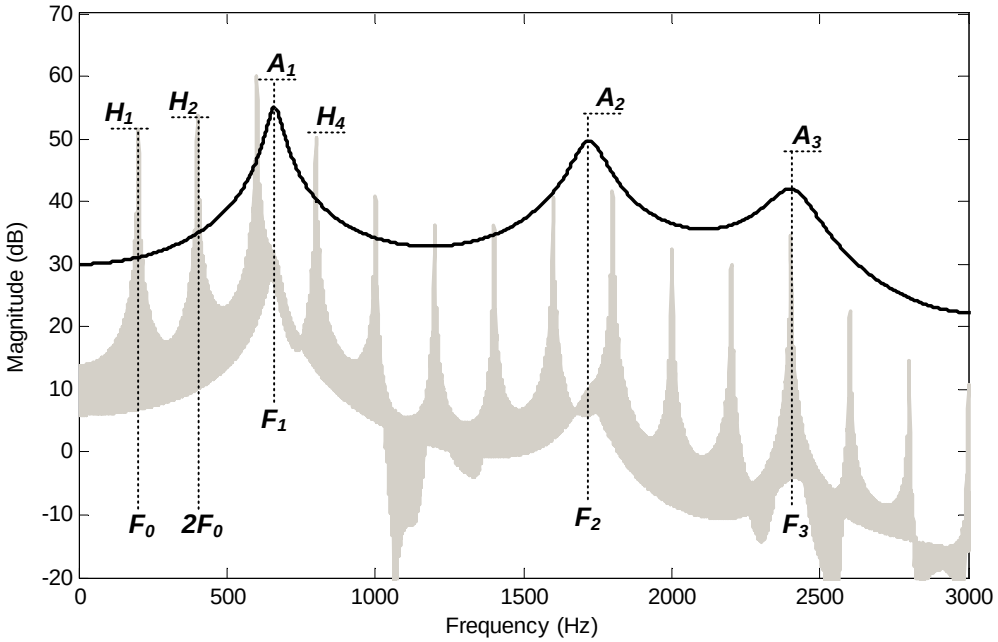



**Figure 1.9:** Magnitude spectrum of the vowel /æ/ showing the measures: pitch frequency ($F_0$), harmonic amplitudes ($H_1$, $H_2$ and $H_4$), formant frequencies ($F_1$, $F_2$ and $F_3$), and the spectral magnitudes at the formant frequencies ($A_1$, $A_2$ and $A_3$).

**Table 1.1:** Some common voice quality related measures.

| **Measure** | **(Hypothesized) relation to the voice source or voice quality** |
|---|---|
| $F_0$ | Correlated with pitch accents, boundaries, lexical tones and stress. |
| $H_1^* - H_2^*$ | Thought to be correlated with breathiness, and with open quotient ($OQ$, the proportion of time the vocal folds are open during phonation). |
| $H_1^* - A_3^*$ | Thought to be correlated with spectral tilt and hence, the rate of closure of the vocal folds. |
| *Energy* | Related to loudness and voice intensity. |
| *CPP* | Cepstral peak prominence: thought to be correlated with modality and breathiness. |
| *Noise* | Noise in the spectrum can be associated with aspiration noise, which is usually associated with breathiness. |

low normal) nature and are sometimes used by English speakers to signal stress or emphasis. In English, pitch is also used to denote boundaries; for example, a lowering of the pitch towards the end of a sentence is usually used to denote a statement, while a rising of the pitch can be used to denote a question.

$H_1^* - H_2^*$, the difference between the first two harmonic magnitudes corrected for the effects of the VTTF, has often been taken as a correlate of the open quotient ($OQ$), which is broadly defined as the proportion of time the vocal folds are open during a phonation cycle. This relationship can be shown theoretically using a simple sinusoid, but the relationship is more complex for speech sounds. Assuming all other influences are constant, a longer opening of the vocal folds means that the open-phase becomes more closely matched with the pitch period, leading to a stronger fundamental component ($H_1^*$) in the signal spectrum. In [HHP95],

$OQ$ estimates from airflow and EGG recordings from 15 female subjects were used to show that the correlation between $H_1 - H_2$ (uncorrected for VTTF) and $OQ$ was moderate, with $r^2 = 0.48$ and $r^2 = 0.21$ respectively. Since $OQ$ is often thought to be correlated with breathiness ([Huf87, Fis67, SL90]), by association, $H_1^* - H_2^*$ has also been used as a measure of breathiness. In perceptual studies, [HCE94] and [KK90] found $H_1 - H_2$ to be moderately correlated with perceived breathiness. Other studies involving phonation type languages such as Mazatec ([Bla97]), Zapotec ([Ave04]), Khmer ([WJ03]), Gujarati ([Kha09, Esp06]) and Hmong ([EPY09]) have also found that $H_1 - H_2$ can be used to distinguish breathy phonations from non-breathy phonations. More recent studies ([HdD01]) have showed that the relationship between $H_1^* - H_2^*$ and $OQ$ is not as strong as previously thought and depends on other voice source parameters such as the asymmetry coefficient (proportion of opening phase duration to closing phase duration).

$H_1^* - A_3^*$, the difference between the VTTF-corrected first harmonic magnitude and the corrected spectrum level at the frequency of the third formant, was shown in [Han97] to be related with the source spectral tilt. Source spectral tilt or spectral balance typically measures the amount of high frequency components relative to low frequencies. They have been used in many voice quality studies ([SV96b, SV96a, CHC05, ISE06]) as a correlate of stress and intonation. It is generally hypothesized that words with more stress or emphasis will lead to tenser vocal folds which contain more high spectral frequency components during phonation.

Energy, unlike other voice source related measures, is known to be correlated with loudness and voice intensity. However, there have been many measures used to represent energy such as the amplitude ([CHC05]), the $E_e$ parameter of the LF

model ([ISE06, Eps02]), and the energy content from specific frequency sub-bands ([RH06]). In terms of voice quality, energy has been shown to be a good correlate of pitch accents ([CHC05, RH06]) and intonational boundaries ([CHC05, Sli07]), although it was found in [ISE06] that it was important to differentiate between the high and low tones in pitch accents due to the relationship between $F_0$ and energy.

*CPP*, the cepstral peak prominence, is defined in [HCE94] as "a measure of cepstral peak amplitude normalized for overall amplitude". In theory, the peaks in the cepstral domain (conventionally known as "rahmonics") reflect the properties of the source, and a well defined periodic source should have larger peaks than a less periodic one. Hence, the *CPP* value should be larger for modal phonations and smaller for breathy phonations which have more noise in the cepstral domain; it can also be smaller for creaky voices if the phonation is aperiodic.

Noise in the speech spectrum is usually thought to be correlated with breathiness. In [KK90], perceptual experiments were used to show that when random noise was added to a synthesized source signal with a large $H_1 - H_2$, English listeners were more likely to rate the signal as being breathy than if only $H_1 - H_2$ was used by itself. However, other studies ([Fis67, Bic82]) have shown inconsistent results regarding the importance of noise for perceiving breathiness.

Other measures have also been used such as $H_1^* - A_1^*$, $H_1^* - A_2^*$ and $H_2^* - H_4^*$. However, as with the measures in Table 1.1, the relationship between these parameters and perceived voice quality have not been extensively studied. In [KGB07], 78 voice source related measures were calculated for the vowel /a/, and analyzed together with synthesized voice source pulses to determine which measures captured the majority of the information in a given pulse. The results showed that

there were many overlaps in what each measure captured and there were also difficulties with modeling the higher frequency parts of the source spectrum. The inconsistencies between the various studies that use voice source related measures can be attributed largely to the lack of empirical evidence relating these measures with the actual voice source, i.e. the movements of the vocal folds.

### 1.4.2 Prosody

Prosody, of which voice quality is a part, is a term which is used broadly to refer to the intonation, rhythm, timing, phrasing and stress in speech. In connected speech, prosody serves both as a grouping function and a prominence-marking function. The groups can be phrases or sentences and are indicated by prosodic *boundaries*, e.g. the delay or break between the phrases or sentences. In English, the prominence of a word within a phrase is marked by particular $F_0$ patterns, called *pitch accents*; e.g. a pitch accent can signal a focal accent, for contrastive stress on a word.

Perceptually, prosodic events can help speakers emphasize particularly important parts of speech, distinguish between word meanings and signify the conclusion of sentences or questions. Likewise, for listeners, the prosodic features of speech can also help to signal the timing of turn-taking in conversational speech. Previous studies of prosody have mainly focused on $F_0$, duration and intensity as acoustic correlates. Because voice quality information is mainly carried in the voice source, it can be expected that voice source related measures would help in the study of prosody. In [CHC05], the harmonic structure and spectral tilt were used with pitch, duration and amplitude information to perform automatic accent and boundary detection experiments on a large prosodically-labeled corpus. Detection rates of approximately 70% were obtained for both accent and

boundary detection. However, in that study, the different types of accents (high vs. low) and boundaries (high vs. low) were not differentiated. In [ISE06] and [Ise07], it was found that low and high types of pitch accents and boundaries were correlated differently to the voice source related measures. In this research, the effects of multiple prosodic events on the measures are studied.

## 1.5 Dissertation outline

This dissertation is organized as follows.

Chapter 1 presented a brief summary of the human speech production process, including the voice source, vocal tract and the voice quality aspects.

Chapter 2 proposes a new model of the voice source based on the glottal area waveforms obtained from the high-speed imaging of the vocal folds.

Chapter 3 introduces a codebook search technique for estimating the voice source.

A new software package is presented in Chapter 4 which simplifies the calculation of voice source related measures. The software is then used in three applications: voice source analysis, gender detection and prosody analysis.

Chapter 5 looks deeper into the effects of multiple prosodic targets on the voice source related measures: $F_0$, energy and duration.

Finally, Chapter 6 summarizes this dissertation and discusses future research directions.

# CHAPTER 2

# A New Voice Source Model based on High-Speed Imaging of the Vocal Folds

Voice source models have been in existence for many decades. A model allows a variety of source configurations to be compactly represented by a few parameters. A good source model can improve the naturalness of synthesized speech. In speech analysis, a detailed source model can help to capture the important properties of a particular speaker. Existing source models have typically been based on data coming from indirect source observations. In this chapter, high-speed video recordings of vibrating vocal folds were processed to produce glottal area waveforms. From these waveforms, a new source model is proposed and evaluated.

## 2.1 Existing voice source models

Existing voice source models were reviewed in Section 1.2.1. Most of these sources such as the Fant, Ananthapadmanabha and LF models, were based on indirect voice source measurements such as the inverse-filtering of speech signals or inverse-filtered airflow measurements. The Rosenberg models (polynomial and trigonometric) were obtained by analysis-by-synthesis experiments.

In this chapter, the LF model ([FLL85]) is used as a basis for the derivation of

a new source model. Since its inception in 1985, the LF model has been the most widely-used model for voice analysis due to its flexibility. It is also currently the most commonly-used source model in speech synthesizers. Figure 2.1 shows the parameters of the LF model, and the equation defining this model is Eq. 2.1. In this equation, the parameters $E_0$, $\alpha$ and $\epsilon$ denote the amplitude scaling factor, growth factor and the exponential time constant of the return phase, respectively. Like the Ananthapadmanabha and Fujisaki-Ljungqvist models, the LF model specifies the derivative flow and not the actual flow. In [FLL85], the LF model was shown to provide a better fit than the Ananthapadmanabha model to inverse-filtering data obtained from one adult male speaker of Swedish. The authors of that study also suggested that the LF model could also be used to describe glottal area waveforms, although this claim has not been verified.

$$u(t) = \begin{cases} E_o e^{\alpha t} \sin(\omega_g t), & 0 \leq t \leq t_e \\ \left(\frac{-E_e}{\epsilon T_a}\right) \left[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}\right], & t_e < t \leq t_c \end{cases} \tag{2.1}$$

Although existing source models have been used with varying degrees of success in voice synthesis applications, the limitations of these models become apparent when used for voice analysis. Utilizing inverse-filtering of singing and speech signals, [HdD01] showed that many source spectra were beyond the scope of what could be generated with existing source models. In that study, it was concluded that this could be due to either the deficiencies of current models or the inaccuracy of inverse-filtering because of source/tract interactions.

Current source models have mainly been derived from some form of empirical data, typically from indirect sources such as inverse-filtered flow measurements or EGG observations. However, without direct observations of the physical source, it is difficult to quantify the accuracy of a particular model.

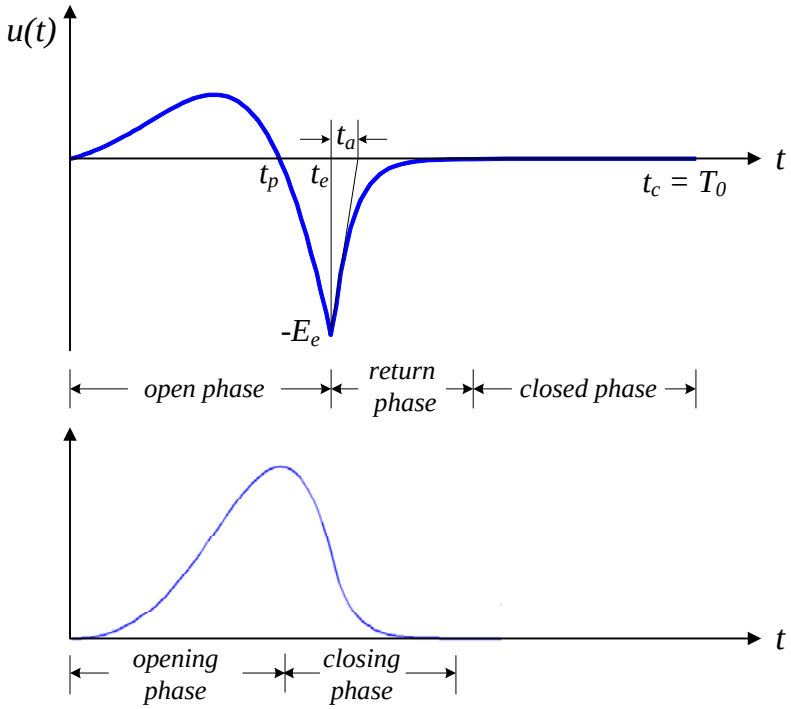



**Figure 2.1:** The LF model. Top panel illustrates the glottal flow derivative: instant of maximum airflow ($t_p$), instant of maximum airflow derivative ($t_e$), effective duration of return phase ($t_a$), beginning of closed phase ($t_c$), fundamental period $T_0$, and amplitude of maximum excitation of glottal flow derivative ($E_e$). Bottom panel illustrates the glottal flow model.

## 2.2 High-speed imaging of the vocal folds with synchronous audio recordings

Synchronous audio and high-speed video recordings were performed on six subjects, three females (denoted by FM1–3) and three males (denoted by M1–3), all with perceptually normal voices. These recordings were performed at the Head and Neck Surgery Division of the UCLA Medical School under the supervision of Professor Jody Kreiman. The high-speed imaging was performed on the vocal folds at 3000 frames/second at a resolution of 512 $\times$ 512 pixels using a 70° rigid laryngoscope with a 300 watt Xenon light source (Kay Elemetrics) and a FASTCAM-ultima APX camera (Photron Ltd., San Diego). Audio recordings were synchronously transduced with a Brüel & Kjær microphone (1.27 cm diameter; type 4193-L-004) and directly digitized at a sampling rate of 60 kHz (conditioning amplifier: NEXUS 2690, Brüel & Kjær, Denmark; bandpass filtering of microphone signal between 20 Hz and 22.4 kHz; AD converter: voltage resolution 16 bits, input range $\pm$ 5 volts). The audio recordings were further downsampled to 16 kHz for analysis.

### 2.2.1 Subjects and voice samples

The six subjects were asked to sustain the vowel /i/ for approximately 10 seconds while holding voice quality, $F_0$, and loudness as steady as possible. The use of the vowel /i/ results in an anterior retraction of the epiglottis, thereby providing the most complete visualization of the vocal folds during phonation [DBP07]. Although the subjects were asked to pronounce the vowel /i/, the vowel quality ranged from /I/ to approximately /æ/ due to the positioning of the laryngoscope over the tongue. During the recordings, subjects were also directed to vary $F_0$

(low, normal and high) and voice qualities (pressed, normal and breathy) quasi-orthogonally, resulting in a minimum of nine recordings from each speaker. One second samples of the phonations were extracted from the most stable and representative portions of each recording for subsequent analysis. The mean $F_0$ values for the six speakers are shown in Table 2.1. For $F_0$ values in the 100–400 Hz range, and images recorded at 3000 fps, there were approximately 7–30 frames per cycle of phonation. Although the six subjects were directed to produce a total of nine different phonations, one female, FM2, was not able to produce a pressed phonation with normal $F_0$, and one male, M1, was unable to produce a low $F_0$ for the three voice qualities.

**Table 2.1:** Mean $F_0$ values for the low, normal and high $F_0$ with pressed (p), normal (n) and breathy (b) voice qualities. '–' denotes data was not available for a particular phonation.

| | mean $F_0$ values (Hz) for the different phonations | | |
|---|---|---|---|
| Subject | low $F_0$ (p/n/b) | normal $F_0$ (p/n/b) | high $F_0$ (p/n/b) |
| FM1 | 177/152/145 | 237/210/198 | 374/336/301 |
| FM2 | 218/168/163 | –/211/214 | 288/341/242 |
| FM3 | 157/196/188 | 248/219/244 | 366/426/307 |
| M1 | –/–/– | 147/137/135 | 259/230/171 |
| M2 | 125/103/99 | 196/129/173 | 289/204/213 |
| M3 | 93/91/98 | 140/135/132 | 223/201/202 |

### 2.2.2 Image segmentation

The glottal area estimation method is shown in Figure 2.2. At the start of each new phonation, an area of interest is manually selected from the 512 × 512 image. This area, shown by the bounding box in the figure, is different for

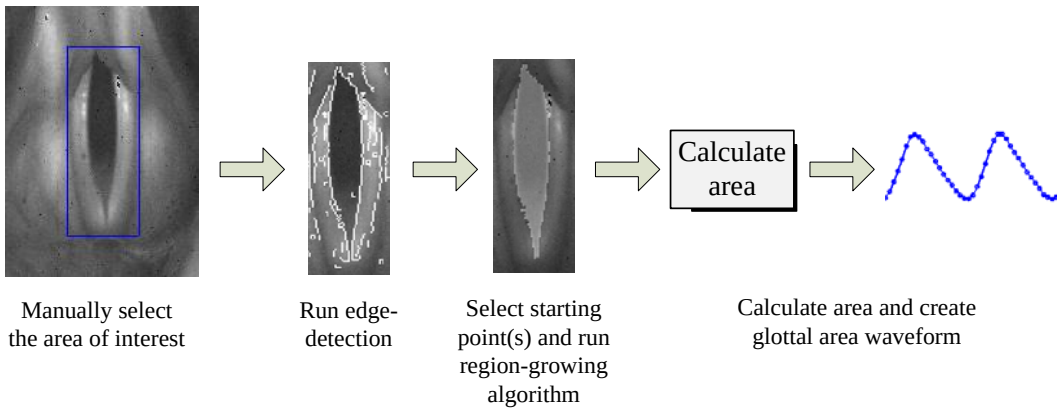



**Figure 2.2:** Glottal area estimation procedure. Images shown are from the low $F_0$, breathy phonation of subject FM1.

each phonation and depends on the distance of the camera from the vocal folds. An edge-detection algorithm with adjustable threshold is then applied to the selected area to produce a segmented image. This threshold can be manually adjusted, if required, after visual inspections of the results. From this segmented image, a starting point for the region-growing algorithm is found automatically by seeking out the darkest pixel within the segmented image. This starting point can be manually overridden if (1) there is much noise in the image, or (2) the vocal fold openings contain more than one glottal gap. Region-growing on the starting point(s) is then performed horizontally on each row of the image, using boundaries from the segmented image to ensure that no pixels outside of the vocal fold opening are selected. The number of pixels selected for each row is constrained to be no greater than a certain threshold from the previous row's selected pixels. This ensures that gaps in the edge-detection do not lead to erroneous pixels being counted. The glottal area is then calculated by adding the number of pixels selected in total.

### 2.2.3 Glottal area estimation

The glottal area estimation process was performed using custom-written software. For images with good contrast levels, as was the case with a few phonations, it was possible to run the entire procedure automatically. However, due to various factors such as random noises, over-exposures, image contrast levels, and multiple glottal gaps, manual adjustments were required for most images.

To reduce processing time, only the first 150 frames (equal to 50 ms) of each utterance were segmented to extract the glottal areas, resulting in 3 to 15 glottal cycles depending on the $F_0$ used in a particular phonation. Each cycle of glottal vibration was then marked by recording, where it existed, the first instances of glottal opening. In samples where there were no complete glottal closures, the least glottal opening instances were recorded. These points allowed for the measurements of the glottal area waveforms to be averaged across glottal cycles to produce a waveform which is representative of the 150 analyzed frames for a particular phonation. The averaging procedure also helped smooth out noise from the segmentation process. For comparison, the averaged glottal area waveforms were then time and amplitude normalized to a length of 100 samples and a maximum height of 1. Time normalization was done using resampling and, where they existed, the extra ripples from the resampling process were removed by zeroing out the values past the main glottal pulse. An example of the averaging and normalization process is shown in Figure 2.3.

The nine averaged glottal area waveforms for subject FM1 are shown in Figure 2.4. The waveforms for the other subjects can be seen in Appendix A. From these waveforms, a few interesting properties can be observed. First, it was noticed that in many cases the opening phase duration was shorter than the closing phase duration. This is not accounted for in the LF model. Second, in some

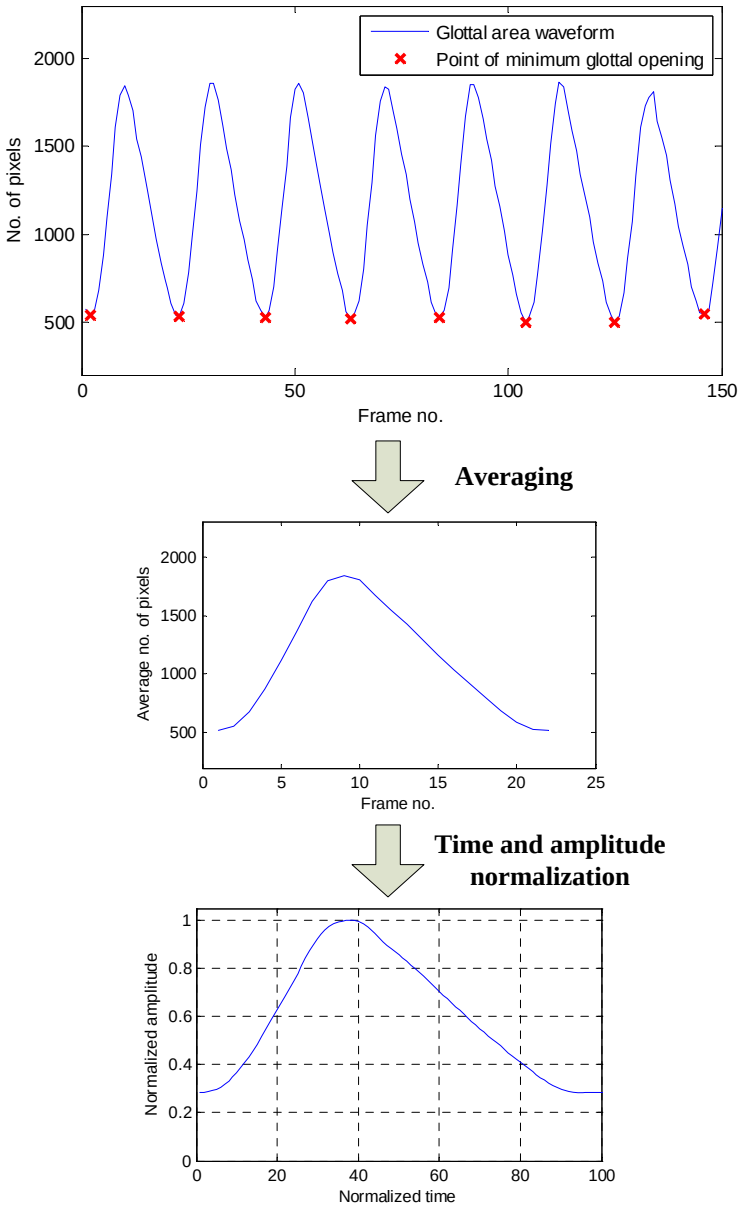



**Figure 2.3:** Glottal area waveform averaging and normalization. Waveforms are from the low $F_0$, breathy phonation of subject FM1.

cases, mostly involving pressed phonations, both the opening and closing phases can occur very quickly, much quicker than what can be specified by the LF model. Finally, the waveforms showed that there should be more flexibility with the functions describing the opening and closing phases. Since the LF model specifies the flow derivative, its functions do not specify the opening and closing phases, but the open and return phases.

## 2.3 A new voice source model

Using the new glottal area data, a new source was proposed to account for properties described in the previous section. In the LF model (Eq. 2.1), a combination of exponential functions are used for the return phase. This is problematic because the decaying property of the exponential does not allow for a quick return to zero. A combination of an exponential function with a sine function, similar to the first equation of the LF model, was determined to be a better fit. Other functions were also tested, but they were either too complex or did not have the required properties to match the glottal area waveforms.

The proposed model builds upon the first equation of the LF model; since the LF model specifies the derivative functions, the integrated version of the first equation is needed to describe the area waveforms. Using the notation from Eq. 2.1, this can be derived (using integration-by-parts or standard integration tables) as:

$$\begin{aligned} U(t) &= \int_0^t E_o e^{\alpha\tau} \sin(\omega_g \tau) d\tau \\ &= \frac{E_o e^{\alpha t}}{\alpha^2 + \omega_g^2} \left[\alpha \sin(\omega_g t) - \omega_g \cos(\omega_g t) + C\right] \end{aligned} \tag{2.2}$$

where $C$ is some arbitrary constant of integration. With the initial condition

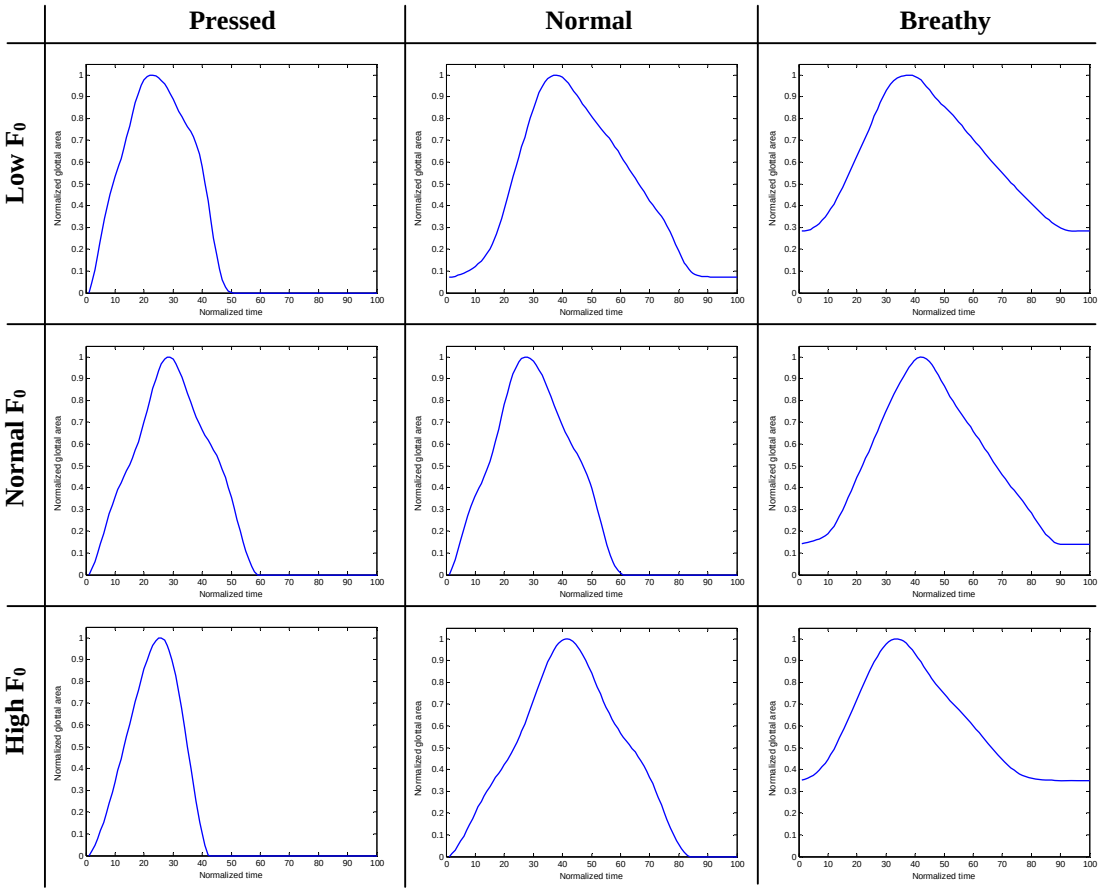



**Figure 2.4:** The averaged glottal waveforms for the nine phonation combinations for subject FM1. $F_0$ (low, normal and high) was varied quasi-orthogonally with voice quality (pressed, normal and breathy). Note that the three voice quality types differ little in the opening, with most of the difference seen in the closing and in the minimum values.

$U(t=0)=0$, $C$ can be shown to equal $\omega_g$. Hence, Eq. 2.2 can be written as:

$$U(t)=\frac{E_o e^{\alpha t}}{\alpha^2+\omega_g^2}\left[\alpha\sin(\omega_g t)-\omega_g\cos(\omega_g t)+\omega_g\right] \tag{2.3}$$

This equation forms the basis of the proposed model and is used for both the opening and closing phases.

The proposed model is a time-domain glottal area waveform and consists of 4 parameters: open quotient ($OQ$), asymmetry coefficient ($\alpha$), speed of opening phase ($S_{op}$), and speed of closing phase ($S_{cp}$). An example of a model waveform is show in Figure 2.5, where $T_0$ denotes the pitch period. Using the notation from this figure, $OQ=\frac{t_o+t_c}{T_0}$, $\alpha=\frac{t_o}{t_o+t_c}$, $S_{op}=\frac{t_{oh}}{t_o}$ and $S_{cp}=1-\frac{t_{ch}}{t_c}$ where $t_{ch}$ and $t_{oh}$ are at 50% of the maximum amplitude. OQ specifies the proportion of time the vocal folds are open, $\alpha$ controls the proportion of OQ which is used for the opening phase, and $S_{op}$ and $S_{cp}$ specifies the proportion of time required to reach 50% of the maximum amplitude for the opening and closing phases. The four other parameters all range from 0 to 1. By modifying Eq. 2.3 as:

$$f(x,\lambda^*)=A(\lambda^*)\left[e^{\lambda^* x}(\lambda^*\sin(\pi x)-\pi\cos(\pi x))+\pi\right] \tag{2.4}$$

where $A(\lambda^*)=\frac{1}{\pi(e^{\lambda^*}+1)}$ and

$$\lambda^*=\arg\min_{\lambda}\left|\frac{e^{\lambda s}(\lambda\sin(\pi s)-\pi\cos(\pi s))}{\pi(e^{\lambda}+1)}+\frac{1}{e^{\lambda}+1}-\frac{1}{2}\right| \tag{2.5}$$

with $s=S_{op}$ or $S_{cp}$, the proposed model can be defined as:

$$u(t)=\begin{cases} f(\beta_o t,\lambda_{S_{op}}), & 0\le t\le t_o \\ f(\beta_c(t_o+t_c-t),\lambda_{S_{cp}}), & t_o<t\le t_o+t_c \\ 0, & t_o+t_c<t\le T_0 \end{cases} \tag{2.6}$$

where $\beta_o=\frac{1}{t_o}$, $\beta_c=\frac{1}{t_c}$, and $\lambda_{S_{op}}$ and $\lambda_{S_{cp}}$ are the $\lambda^*$ values when $s=S_{op}$ and $S_{cp}$ in Eq. 2.5, respectively. $A(\lambda^*)$ is a normalizing term so that $\max f(x,\lambda^*)=1$.

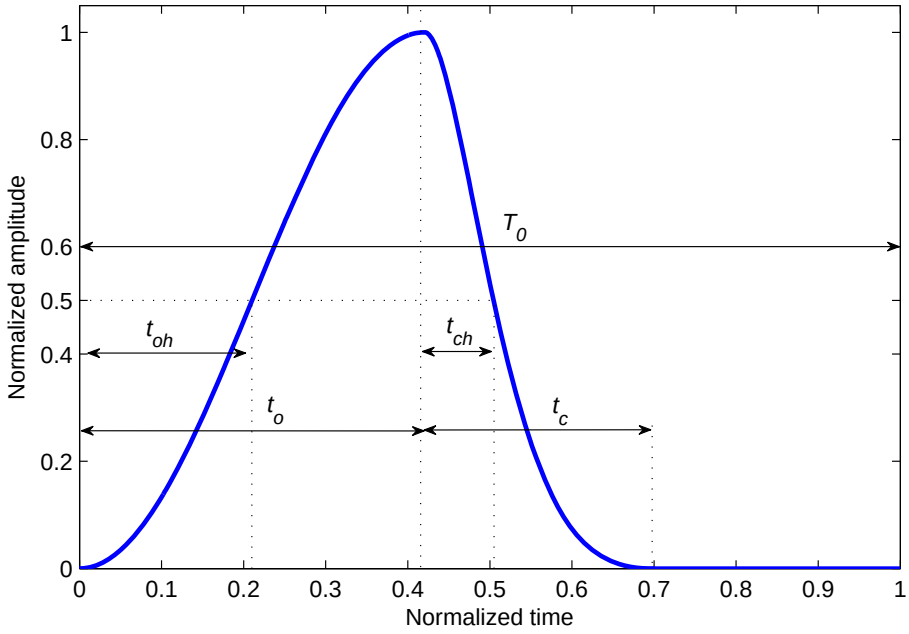



**Figure 2.5:** Example of the proposed source model with $OQ = 0.7$, $\alpha = 0.6$, $S_{op} = 0.5$ and $S_{cp} = 0.7$.

The slope of $f(x, \lambda^*)$ is determined by $\lambda^*$, which can be calculated by simple optimization techniques, such as the gradient descent algorithm. A somewhat non-trivial closed-form solution for $\lambda^*$ also exists for some $s$ values involving the Lambert $W$-function.

### 2.3.1 Properties of the new source model

The proposed source model utilizes the integrated version of the first equation of the LF model to form both the opening and closing phases. This equation allows for quicker transitions between the pulse onset to the pulse peak and also between the pulse peak and the pulse offset. Having equations which specify these two phases separately, like in the Rosenberg model, is important in regards to the new high-speed imaging data because, as shown by the averaged glottal area waveforms, the opening phase can be very different compared to the closing phase.

Unlike the LF and Fujisaki-Ljungqvist models, almost all combinations of parameters of the proposed source model result in a continuous waveform, although some degenerate cases exist. For analysis applications, such as model fitting, this property allows for the parameters to be found quickly.

### 2.3.2 Incomplete glottal closures and the DC-offset

From the waveforms in Figure 2.4 and Appendix A, it can be seen that in some phonations, especially the breathy ones, the glottal folds do not close completely during the closing phase. This leads to a vertical shift in the position of the glottal area waveforms. While this could be modeled by adding a "DC-offset" parameter, it was not done here for two main reasons. First, in the linear speech production model ([Fan70]), lip radiation is often modeled by applying a derivative operation

to the glottal flow signal (see Section 1.2) which would remove the effects of any DC-offset parameter. Secondly, it is not well understood how glottal gaps affect perceived voice quality. Early studies ([Fis67], [KK90] and [SL90]) suggested that glottal gaps are typically perceived as turbulent noise which is usually manifested in the frequency domain as a rise in the noise floor of the speech spectrum. However, empirical evidence to support this hypothesis is difficult to obtain due to the invasive nature of direct observations of the vocal folds. In [Han97], it was hypothesized that large posterior glottal openings may result in larger spectral tilt measures ($H_1^* - A_1$ and $H_1^* - A_3^*$). Her preliminary fiberscopy experiments on four of those speakers showed that the two speakers with large posterior openings did indeed have larger spectral tilt measures than those with minor openings. However, in that study, the fiberscopic image recordings occurred after the audio recordings and as stated by the author, simultaneous recordings are needed to verify the hypothesis. In Section 4.3, the spectral correlates of voice quality in relation to the new high-speed imaging data are examined in more detail.

### 2.3.3 Glottal area or glottal flow?

By definition, the glottal flow, as modeled by the voice source models reviewed in Section 1.2.1, measures the volume velocity of the air produced at the glottis. The glottal area is the area of separation between the vocal folds as projected by the image of the glottis. Although the exact relationship between the glottal area and the glottal flow is still unclear, it is generally thought to be of a nonlinear nature. Studies dealing with the relationship, such as in [AF82], typically use impedance networks to model the effects of the glottal area on the glottal flow. However, without direct empirical evidence, the results of the impedance network experiments remain unverified. In a more recent study ([HM07]), which used

computationally intensive fluid modeling, it was found that while the acoustic source pulse shapes differed from the glottal area waveforms, the differences were small relative to the larger differences across the waveforms. Thus, it is reasonable to compare models of glottal area (like the new proposed source model) to models of glottal flow (like the Rosenberg and LF models).

## 2.4 Evaluation of the new source model

Evaluation of the proposed new source model was done by fitting the model to the high-speed imaging data. The averaged glottal area waveforms obtained in Section 2.2.3 were first normalized to have a minimum value of 0 (i.e. with DC-offset removed) and a maximum value of 1. For each phonation by each subject, the proposed model was fitted to the normalized source pulses by using a mean squared error (MSE) criterion. For comparison purposes, the LF model was also fitted to the glottal area waveforms for each subject. Table 2.2 shows the MSEs for each subject's phonations. As expected, visual inspections (see Appendix B for the model fitting performance of the proposed new model) showed that the proposed new source model was able to provide a better fit to the glottal area waveforms in all cases. This was not surprising given that the proposed model was derived from these very same glottal area waveforms while the LF model was derived from a different set of data (inverse-filtered flow recordings). More interesting are the cases where the MSEs differ the most; two such cases are shown in Figure 2.6. The left panel is from the low $F_0$, pressed phonation from subject FM1 and the right panel is from the low $F_0$, normal phonation from subject M2. It can be seen in these two cases, that the pulse peak is skewed towards the glottal onset, resulting in the opening phase duration being shorter than the closing phase duration. At the pulse onsets and offsets, the quick transitions to

the zero amplitude are unable to be modeled by the slower moving LF model. For comparison, an example of a case where both the proposed and LF models had approximately equivalent MSEs is shown in Figure 2.7. This example is from the low $F_0$, normal phonation from subject FM3. In this case, the pulse peak is approximately at the center and the pulse onset and offset consist of more gradual changes.

**Table 2.2:** Model fitting results for each phonation from each speaker. Results are shown for the proposed/LF models. '–' denotes data was not available for a particular phonation.

| Voice quality | | Subjects (MSE of proposed/LF model) | | | | | |
|---|---|---|---|---|---|---|---|
| | | FM1 | FM2 | FM3 | M1 | M2 | M3 |
| low $F_0$ | pressed | .002/.026 | .000/.005 | .001/.003 | –/– | .001/.005 | .002/.014 |
| | normal | .001/.010 | .000/.006 | .001/.001 | –/– | .000/.035 | .005/.020 |
| | breathy | .000/.017 | .001/.013 | .000/.008 | –/– | .001/.027 | .002/.021 |
| normal $F_0$ | pressed | .002/.012 | –/– | .004/.008 | .001/.011 | .001/.015 | .003/.012 |
| | normal | .002/.015 | .001/.008 | .001/.006 | .001/.014 | .001/.006 | .002/.005 |
| | breathy | .001/.006 | .001/.008 | .001/.001 | .001/.017 | .001/.025 | .001/.004 |
| high $F_0$ | pressed | .000/.003 | .000/.001 | .001/.013 | .000/.002 | .000/.006 | .001/.006 |
| | normal | .003/.010 | .002/.008 | .004/.010 | .000/.007 | .005/.023 | .000/.002 |
| | breathy | .000/.010 | .000/.016 | .001/.004 | .000/.023 | .000/.029 | .000/.025 |

The proposed new source model will also be evaluated in Chapter 3 by a source estimation technique which used a codebook of source signals. Comparisons with the LF model will show that in most cases, the proposed new model provided a more accurate estimate of the source signal.

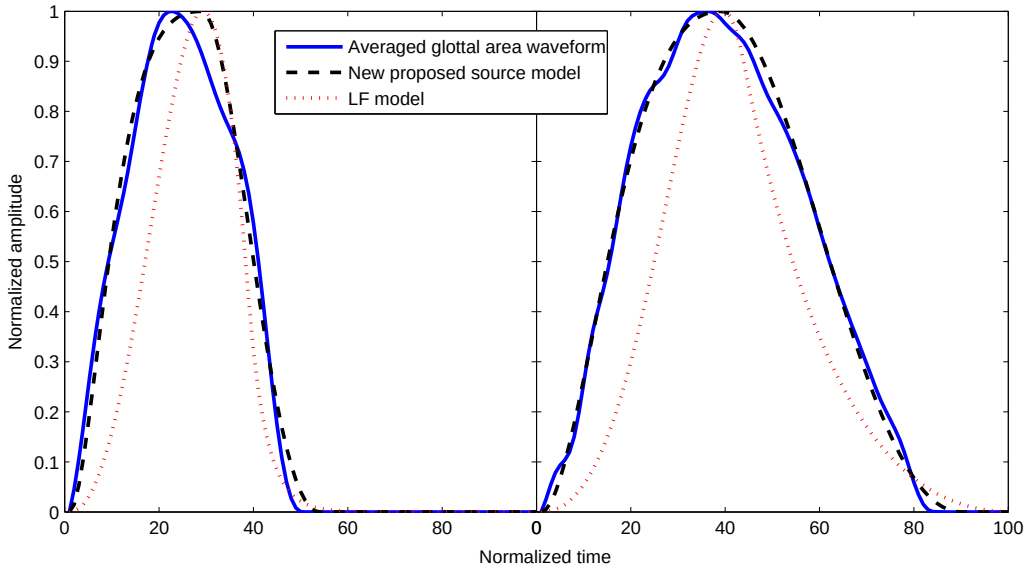



**Figure 2.6:** Model fitting performance for the phonations from subject FM1 (left panel: low $F_0$, pressed) and subject M2 (right panel: low $F_0$, normal).

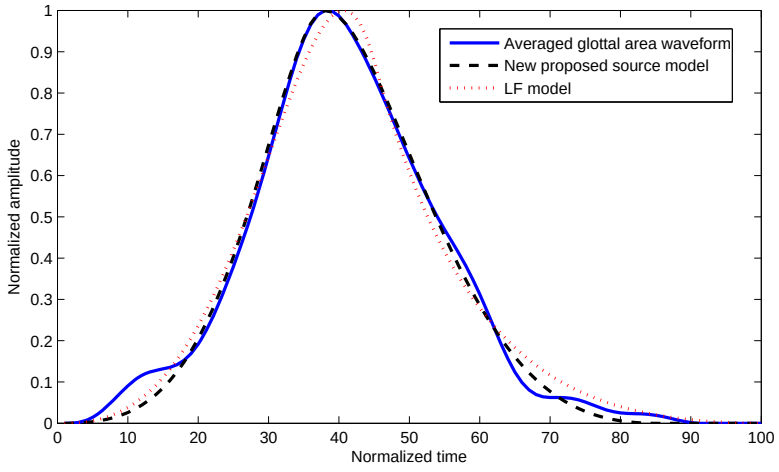



**Figure 2.7:** Model fitting performance for the low $F_0$, normal phonation from subject FM3.

## 2.5 Summary and discussion

In this chapter, high-speed video recordings of the glottal folds were extracted and analyzed to produce glottal area waveforms. The analysis showed that there were some properties of the area waveforms which could not be modeled by existing source models. By modifying the popular LF model, a new four parameter source model was created which was able to better account for the properties observed in the glottal area waveforms.

The quantitative comparisons between the LF and the new proposed source model are not so much a measure of which model is the most accurate, since the models were derived from different data sets with different purposes, but an indication of how existing source models can be used to motivate the creation of newer models.

In source modeling, many unsolved issues still remain. Previous source models were created based on indirect observations of the source, and as such modeled the glottal flow. Improved technology has allowed easier access to observe the vibrating vocal folds, and the source model proposed in this chapter models the glottal area. The exact relationship between the glottal flow and the glottal area waveform, which has been hypothesized to be non-linear and small in effect, remains an open question. Also unsolved is the ubiquitous glottal gap (DC-offset) issue. While perceptual studies suggest that glottal gaps introduce noise, the exact nature of how the noise is created is not well known. More data from the direct observation of the glottis is needed to resolve these issues.

# CHAPTER 3

# A Codebook Search Technique for Estimating the Voice Source

Voice source estimation is a non-trivial task which requires the separation of the source signal from the speech signal. Traditional methods have employed inverse-filtering or joint estimation techniques to extract the source signal. However, these techniques rely on the assumption that speech production is a linear and time-invariant process, which it is not. The non-linear interactions between the source and the vocal tract can result in inaccuracies which may be reflected in both the source signal and vocal tract filter estimates. In this chapter, an analysis-by synthesis technique is introduced which, different from previous methods, effectively performs inverse-filtering with the source signal spectrum, instead of the vocal tract spectrum. Results are evaluated using direct glottal observations from the high-speed imaging described in Chapter 2.

## 3.1 Voice source estimation

According to the linear acoustic theory of speech production [Fan70], speech signals are generated by a source or excitation signal filtered by the vocal tract transfer function (VTTF). In many applications, we are interested in the underlying acoustic features of the source signal because it can carry information

regarding stress (or emphasis), emotional status, prosodic events, or even an underlying disease of the vocal cords. Estimation of the source signal is a non-trivial process as it requires separating the source from the VTTF. Typical voice source estimation methods (reviewed in Section 1.3) involve the initial estimation of the VTTF, followed by inverse-filtering of the speech signal to obtain a residual signal which is then used to map to a source model. This method is often used in speech coding, where the error criterion is usually based on the re-synthesized output and not on the accuracy of the estimated source or vocal tract. Figure 3.1 shows the typical inverse-filtering process. It can be seen that the resulting residual signal appears more like random noise[1] than the smooth plot shown in Figure 2.1. This is due to a number of reasons: (1) there may be actual aspiration noise in the phonation, although this is generally minimal in modal-type phonations, (2) the process relies heavily on an accurate estimation of the vocal tract filter, and (3) the process enforces the linear model onto the vocal tract, leaving the residual signal to carry the non-linear source-tract interaction information. Another difficulty with source estimation is the lack of a "ground truth" with which to validate algorithms. Often, calibrations are performed with analysis-by-synthesis results, which minimizes the re-synthesized output error, or with EGG signals, which measures the glottal contact area and is only partly related to the voice source signal.

In [IC04], the problem of accurately estimating the VTTF was addressed by using variable window lengths to better capture the vocal tract parameters in the glottal closure regions. The resulting estimated source waveforms were comparable to waveforms obtained from electroglottography (EGG). Joint estimation techniques such as [FMS01], [PY08] and [JI05] attempt to estimate the

[1] In code excited linear prediction (CELP) based speech codecs, Gaussian noise is actually used as the excitation signal.

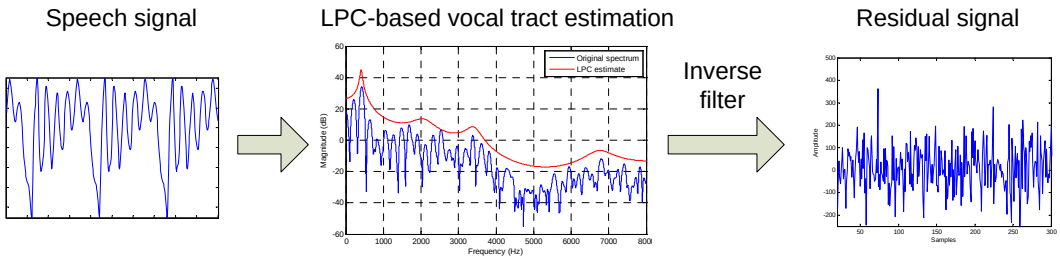



**Figure 3.1:** A typical inverse-filtering process. The residual signal is then used to map onto a source model.

VTTF and voice source model parameters simultaneously. The basic assumption in the two source estimation methods (inverse-filtering and joint estimation) is that speech production can be approximated by a linear time-invariant process. However, it is well known that during speech production, source-tract coupling occurs which can result in non-linear effects. In the inverse-filtering method, these non-linearities usually appear in the residual signal, which is then used for source-model fitting. In the joint-estimation method, the non-linearities may be incorporated into both the source parameters and the VTTF.

In order to minimize errors in the source signal estimation, it is necessary to switch the roles of the source model and the VTTF in the standard inverse-filtering process. In this chapter, a method is proposed in which the voice source is used in inverse-filtering, resulting in a residual signal which is used to fit to the parameters of the VTTF. By reversing the roles of the VTTF and source, it is hoped that the non-linear effects and other noises can be mapped onto the parameters of the VTTF, thereby providing a more accurate source estimate.

## 3.2 Data

The audio data used in this section are the same as those which were recorded synchronously with the high-speed imaging data described in Section 2.2. The measured voice source waveforms used in this chapter are also the same as those described in Section 2.2.3, but with the DC-offset removed as no current time-domain source model has the ability to model this offset.

As in Chapter 2, the male subjects are denoted by M1–3 and the female subjects are denoted by FM1–3.

## 3.3 Method

The linear source-filter model of speech production (reviewed in Section 1.2) states that for short-time periods, speech, $s(t)$, can be approximated as a cascade of linear systems involving a source function, $u(t)$, a vocal-tract transfer function, $v(t)$, and a differentiation which is usually incorporated into the source function:

$$s(t) = u(t) * v(t)$$

$$S(\omega) = U(\omega)V(\omega)$$

Taking the magnitudes of each system, $V(\omega)$ can be written (in dB) as:

$$|V(\omega)| = |S(\omega)| - |U(\omega)| \tag{3.1}$$

Eq. 3.1 implies that if a source spectrum was known, the vocal tract spectrum could be calculated exactly. However, estimation of the spectrum, $|S(\omega)|-|U(\omega)|$, is not robust, and can often result in spurious values near the valleys of $S(\omega)$. More robust are the harmonic magnitudes, as used in [FMS01], denoted by $|S(\omega_{Hk})|$, $|U(\omega_{Hk})|$ and $|V(\omega_{Hk})|$, where $\omega_{Hk} = 2\pi k F_0/F_s$, $k \in \mathbb{Z}^+$. Furthermore, the effects of the overall signal power can be neglected if the harmonic

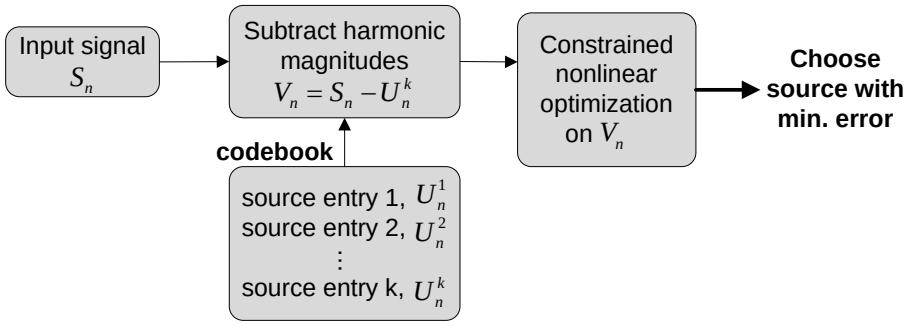



**Figure 3.2:** The main source estimation procedure.

magnitudes are normalized to the first harmonic magnitude; e.g. in dB, $S_n = |S(\omega_{H1})| - |S(\omega_{Hn})|$.

The main block diagram of the method is shown in Figure 3.2. In this method, a codebook of source signals is used to implicitly inverse-filter an input signal, leaving the residual signal for the VTTF and other non-linear source-tract interactions. In the algorithm, the harmonic magnitudes of an input signal are calculated and normalized to the first harmonic magnitude; this is denoted by $S_n$ (for the $n$-th normalized harmonic magnitude) in Figure 3.2. $S_n$ was calculated using the pitch information extracted by the STRAIGHT algorithm [KCP98]. A Hamming window consisting of 4 pitch periods was used to calculate the spectrum of the input signal; hence, the window length was different for each speaker. For this work, the number of harmonics used was in the range 0 to 2.6 kHz; e.g., for a pitch period of 100 Hz, 26 harmonics would be used. This number is arbitrary and in practice, depends only on the number of harmonics that can be reliably estimated from the spectrum.

The codebook of source models can be created using any source model. The block diagram for the codebook generation is shown in Figure 3.3. Grid searches are first performed on the model parameters to find valid voice source waveforms.

This is more critical for models such as the LF or Fujisaki-Ljungqvist models as not all combinations of parameters can be solved to return a continuous signal. The valid source waveforms are then differentiated if necessary (to incorporate lip radiation effects) and used to calculate the harmonic magnitudes. The harmonic magnitudes are normalized to the first harmonic magnitude before being stored in a codebook; these are denoted by $U_2^1, U_3^1, \ldots, U_n^k$ in Figure 3.3. In this chapter, the LF model and the proposed new model in Chapter 2 were used to create two separate codebooks for comparison.

The LF model's codebook was generated by performing a grid search on each of the four parameters ($t_e$, $t_p$, $t_a$ and $E_e$) at the following resolutions: $t_e$ from 0.3 to 0.98 at increments of 0.01, $t_p$ from 0.01 to 0.95 at increments of 0.01, $t_a$ from 0.01 to 0.95 at increments of 0.01 and $E_e$ from 0.1 to 5 at increments of 0.1. Since not every combination of the four parmeters constituted a valid glottal flow derivative waveform, the resulting signals were checked to ensure they were physically realizable. The number of entries in the codebook was reduced by performing a correlation analysis and discarding those entries which had a correlation coefficient of 0.99 or more. This resulted in a final codebook size of 1726 entries.

The proposed new model's codebook was generated with the following parameter resolutions: $OQ$ from 0.35 to 1.00 at increments of 0.01, $\alpha$ from 0.35 to 0.5 at increments of 0.05, $S_{op}$ from 0.3 to 0.7 at increments of 0.2, and $S_{cp}$ from 0.3 to 0.7 at increments of 0.2. This produced a codebook of size 2179 entries after the correlation analysis. The asymmetry coefficient, $\alpha$, only ranged half of its possible values due to the property of the Fourier transform for time-reversed signals; e.g. for a signal $h(t)$ with periodicity $T_0$, $|\mathcal{F}\{h(t)\}| = |\mathcal{F}\{h(T_0 - t)\}|$, where $\mathcal{F}$ denotes the Fourier transform, and a source with $\alpha = 0.4$, $S_{op} = 0.6$

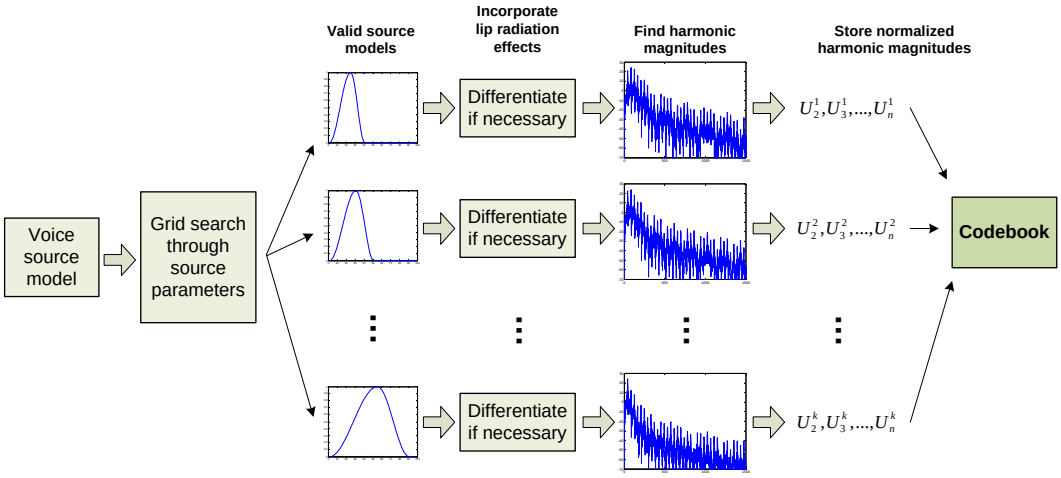



**Figure 3.3:** Block diagram showing the method for generating the codebook. Any voice source model can be used to generate the codebook.

and $S_{cp} = 0.5$ is a time-reversed version of a source with $\alpha = 0.6$, $S_{op} = 0.5$ and $S_{cp} = 0.6$ for any $OQ$ value. While it is not yet clear what perceptual difference, if any, can be noticed between a source and its time-reversed variant, a simple analysis-by-synthesis test, in the time-domain, was employed at the end of the main algorithm to decide which version of the source should be selected.

For each entry in the codebook, denoted by $U_n^k$ for the $k$-th source entry and the $n$-th normalized harmonic magnitude, subtraction with $S_n$ is performed to produce an estimate of the normalized harmonic magnitudes of the vocal tract, i.e. $V_n = S_n - U_n^k$. A constrained nonlinear optimization, using the active-set quadratic programming method, is then performed on $V_n$ to find an estimate of the formant frequencies and their bandwidths as well as an error value. Since the speech data used in this work involved just vowels and the number of harmonics used ranged only up to 2.6 kHz, a 3-formant (6-pole) model was used for the vocal tract:

$$|V(\omega)|^2 = \prod_{p=1}^{3} \frac{1}{\left(1 - 2r_p \cos(\omega - \omega_p) + r_p^2\right)\left(1 - 2r_p \cos(\omega + \omega_p) + r_p^2\right)}$$

where $r_p = e^{-\pi B_p/F_s}$, $\omega_p = 2\pi F_p/F_s$, $F_s$ is the sampling frequency and $F_p$ and $B_p$ are the formant frequencies and their respective bandwidths. Note that, different from other source estimation methods, the optimization here is over the VTTF parameters and not the voice source. The optimization criterion is a weighted least squares error function:

$$E_k = \min_{V_n} \sum_{n=2}^{N} \left(S_n - U_n^k - V_n\right)^2 \cdot W_n$$

where $V_n = |V(2\pi F_0)|^2 - |V(2n\pi F_0)|^2$, $N$ is the number of harmonics up to 2.6 kHz and $W_n$ is a weighting function used to emphasize the lower frequency harmonic magnitudes. For the results presented in this chapter, $W_n$ was empirically defined as:

$$W_n = \begin{cases} 2^{12-n}, & 2 \leq n \leq 12 \\ 1, & n > 12 \end{cases}$$

The constrained optimization on $V_n$ to determine the vocal tract parameters require lower and upper bounds on the formant frequencies and their respective bandwidths. In this work, three methods of determining formant frequency bounds were tested. The first method used the Snack sound toolkit [Sj04] to estimate the formants frequencies with these settings: window length of 25 ms, frame length of 1 ms and pre-emphasis of 0.98. These formant frequencies were then averaged across a subject's phonations. The second method used formant frequencies which were manually extracted from the spectrum of a subject's normal phonation with normal $F_0$. Although the subjects were asked to produce the vowel /i/ for each recording, the end result was always different due to the positioning of the laryngoscope. For 3 male and 2 female subjects, the produced vowels were closer to an /$\varepsilon$/ vowel, while for the other female subject, the resulting vowels were closer to the /æ/ vowel. Because it may impractical to use

manually-derived formant constraints in applications, the third method used constant average formant values for known vowels (/æ/ for subject FM3 and /$\varepsilon$/ for the other subjects), as listed in [PB52], as estimates.

Using the Snack-estimated, manually-extracted and constant-based formant frequency values, the lower and upper bounds for the constrained optimization were set to $\pm$150 Hz from the $F_1$ values obtained by the three methods (Snack/manual/constant-based), $\pm$250 Hz for the $F_2$ values obtained by the three methods, and $\pm$400 Hz for the $F_3$ values obtained by the three methods. Table 3.1 shows the optimization constraints for the formant frequencies in terms of each subject for the three methods; e.g. the Snack-estimated value for $F_1$ for subject FM1 was 351 Hz, therefore the lower optimization constraint was set to 201 Hz and the upper constraint set to 501 Hz. The large differences between the Snack-estimated and manually-extracted $F_1$ values can be attributed to the well-known deficiencies of LPC-based formant trackers for high $F_0$ phonations. As shown in Table 2.1, the average $F_0$ values for high $F_0$ phonations range from 288–426 Hz for females and 201–289 Hz for males. In these cases, the formant-frequencies were often biased towards the harmonic positions, due to the increased spacing between harmonics which effectively lowers the formant "resolution" in the spectrum.

The bandwidth constraints were based on the formant-bandwidth mapping formula in [HM95] since the Snack-estimated bandwidths had large variances.

The algorithm shown in Figure 3.2 requires that all the source models be processed before the final source candidate is selected. For large codebooks, this can be very time-consuming. To reduce the overall processing time, two iterations of the algorithm can be used, as shown in Figure 3.4. In the first iteration, shown with solid lines, a smaller codebook is used to find the approximate source

**Table 3.1:** Optimization constraints for formant frequencies for each subject.

| | Snack-based lower/upper bounds (Hz) | | |
|---|---|---|---|
| Subject | $F_1$ | $F_2$ | $F_3$ |
| FM1 | 201/501 | 1466/1766 | 2116/2916 |
| FM2 | 196/496 | 1331/1831 | 2437/3237 |
| FM3 | 464/764 | 1454/1954 | 2550/3350 |
| M1 | 287/587 | 1433/1933 | 2300/3100 |
| M2 | 229/529 | 1310/1810 | 1999/2799 |
| M3 | 176/476 | 1422/1922 | 2423/3223 |
| | Manual-based lower/upper bounds (Hz) | | |
| Subject | $F_1$ | $F_2$ | $F_3$ |
| FM1 | 450/750 | 1430/1930 | 2115/2915 |
| FM2 | 440/740 | 1650/2150 | 2350/3150 |
| FM3 | 680/980 | 1620/2120 | 2550/3350 |
| M1 | 380/680 | 1550/2050 | 2300/3100 |
| M2 | 410/710 | 1350/1850 | 2000/2800 |
| M3 | 380/680 | 1350/1850 | 1900/2700 |
| | Constant formant-based lower/upper bounds (Hz) | | |
| Subject | $F_1$ | $F_2$ | $F_3$ |
| FM1–2 | 460/760 | 2080/2580 | 2590/3390 |
| FM3 | 700/1000 | 1800/2300 | 2450/3250 |
| M1–3 | 380/680 | 1590/2090 | 2080/2880 |

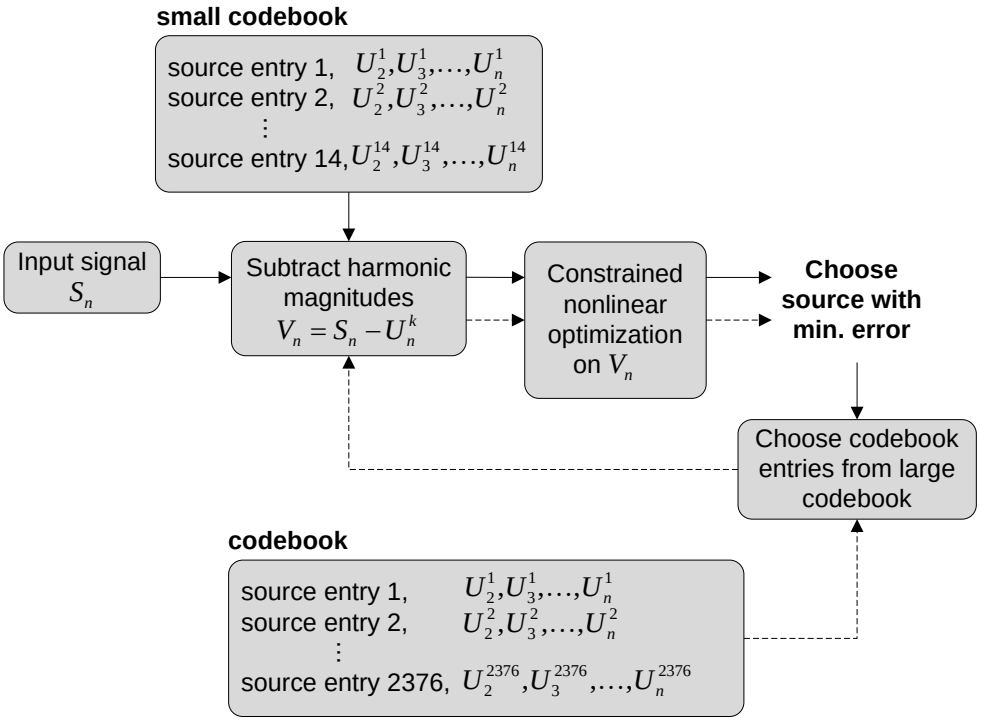



**Figure 3.4:** Block diagram showing the two iterations of the source estimation method; solid and dashed lines represent the first and second iteration, respectively. The codebook sizes are based on the proposed new source model described in Chapter 2.

parameters. These parameters are then used to select source entries, which are within a certain parameter distance, from the main codebook. A second iteration of the algorithm, shown in Figure 3.4 with dotted lines, is then performed using these entries.

In this work, the smaller LF model codebook was created by averaging entries in the main codebook which had $OQ$ values of 0.35, 0.45,. . . ,0.95. This resulted in a codebook with 13 entries. After the first iteration of the algorithm, the source with the smallest error, denoted by $m$, was used to select the entries from the main codebook for the second iteration. Assuming that the $OQ$ for entry $m$ was $OQ_m$, then all entries in the main codebook which had $OQ > OQ_m - 0.1$ and $OQ < OQ_m + 0.1$ were selected for the second iteration. The smaller codebook for the proposed new source model was created with these parameter settings: $OQ$ from 0.4 to 1.0 at increments of 0.1, and $\alpha$ from 0.4 to 0.5 at increments of 0.1. $S_{op}$ and $S_{cp}$ were both set to a constant value of 0.5. This resulted in a codebook of 14 entries. Assuming that the $OQ$ and $\alpha$ values of the selected source entry at the end of the first iteration were denoted by $OQ_m$ and $\alpha_m$, the second iteration used source entries from the main codebook which had an $OQ$ value within $OQ_m \pm 0.1$ and an $\alpha$ value within $\alpha_m \pm 0.05$.

## 3.4 Results

The results of the voice source estimation accuracy were quantified by comparing the shapes of the measured glottal pulses with estimated glottal pulses. The estimated glottal shapes were from the two source models (LF/proposed), with the three different formant frequency constraints (Snack/normal/constant). Table 3.2 shows the mean squared error (MSE) values averaged over a subject's total phonations. It can be seen that nearly all of the averaged MSEs are lower for

**Table 3.2:** Results for each subject and the formant constraint method (Snack, manual and constant based). Values are the MSEs, averaged over all of a subject's phonations, for the LF/proposed new source models.

| | Formant constraint type | | |
|---|---|---|---|
| Subject | Snack | Manual | Constant |
| FM1 | .072/.048 | .091/.035 | .096/.042 |
| FM2 | .105/.049 | .047/.026 | .045/.026 |
| FM3 | .063/.087 | .055/.043 | .047/.025 |
| M1 | .079/.041 | .072/.029 | .072/.029 |
| M2 | .071/.026 | .070/.029 | .085/.028 |
| M3 | .054/.024 | .058/.025 | .058/.025 |

the proposed new source model than the LF model. This is not unexpected due to the deficiencies of the LF model which were discussed in Chapter 2. Another trend which can be seen is that the mean MSEs from the Snack-based formant constraints were on average higher than those from the manual or constant-based formant constraints. These results show the importance of having an accurate source model and also having reasonable formant frequency constraints.

In the rest of this section, the results for the proposed new source model will be analyzed in further detail to see which types of phonations resulted in the lowest and highest estimation errors. The MSE values from the source estimation for individual subjects can be found in Appendix C. With only 6 subjects, there was not enough data to perform statistical analysis.

Visual inspections of the estimated source waveforms (see Appendix C) show that, on average, the source estimation algorithm was able to find the approximate $OQ$ for each phonation, but it could be seen there were some estimation

**Table 3.3:** Correlation coefficients ($r$) for the model-fitted source parameters and the estimated source parameters from the Snack-, manual-, and constant-based formant frequency constraints. The significance levels are in parenthesis, where '–' denotes a particular correlation was not statistically significant.

| Model-fitted | Estimated | | |
|---|---|---|---|
| | Snack-based | Manual-based | Constant-based |
| $OQ$ | 0.616 (0.000) | 0.722 (0.000) | 0.742 (0.000) |
| $\alpha$ | 0.026 (–) | -0.005 (–) | -0.013 (–) |
| $S_{op}$ | 0.220 (–) | 0.369 (0.008) | 0.358 (0.011) |
| $S_{cp}$ | 0.087 (–) | 0.088 (–) | 0.200 (–) |

errors with the other source model parameters ($\alpha$, $S_{op}$ and $S_{cp}$). Using the model-fitted parameters from Section 2.4, correlation analyses were performed on the estimated source parameters for the three types of formant constraints. Table 3.3 shows the correlation coefficients ($r$) and the significance levels (where the correlation was statistically significant). It can be seen that the estimated $OQ$ parameter has the highest correlations, with $S_{op}$ also having some correlation for the manual- and constant-based estimations. The lack of any correlations for $\alpha$ and $S_{cp}$ suggests that changes in these parameters may not be manifested in the harmonic magnitudes.

Figure 3.5 shows the MSE values from the source estimation, averaged over the phonations for each gender in terms of the voice quality and formant constraint type. For the female subjects, it can be seen the Snack-based formant constraints had averaged MSE values which were greater than the other two type of formant constraints. This is not surprising, given the well-known issues associated with LPC-based formant estimation for high-pitched voices. For the male subjects, the averaged estimation errors appear to increase from the pressed phonations

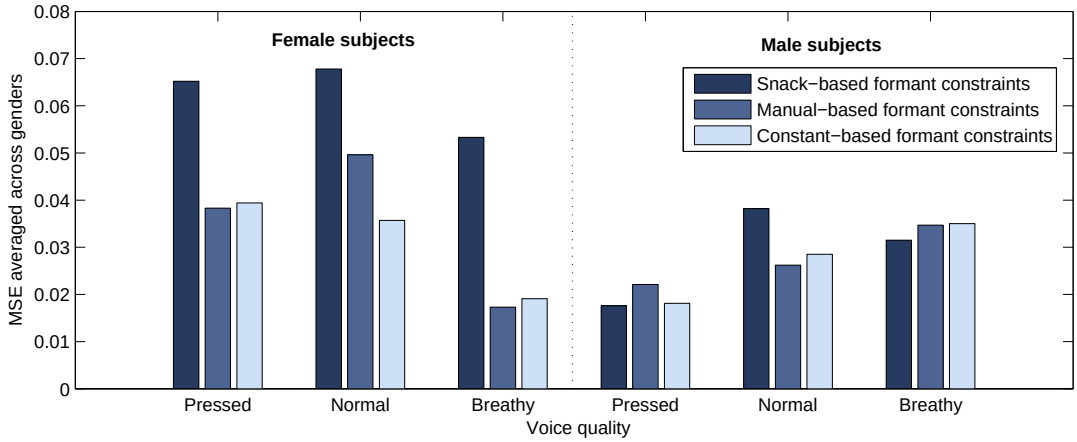



**Figure 3.5:** MSEs averaged across all phonations for each gender in terms of the voice quality (pressed, normal and breathy) and type of formant constraint (Snack, manual and constant).

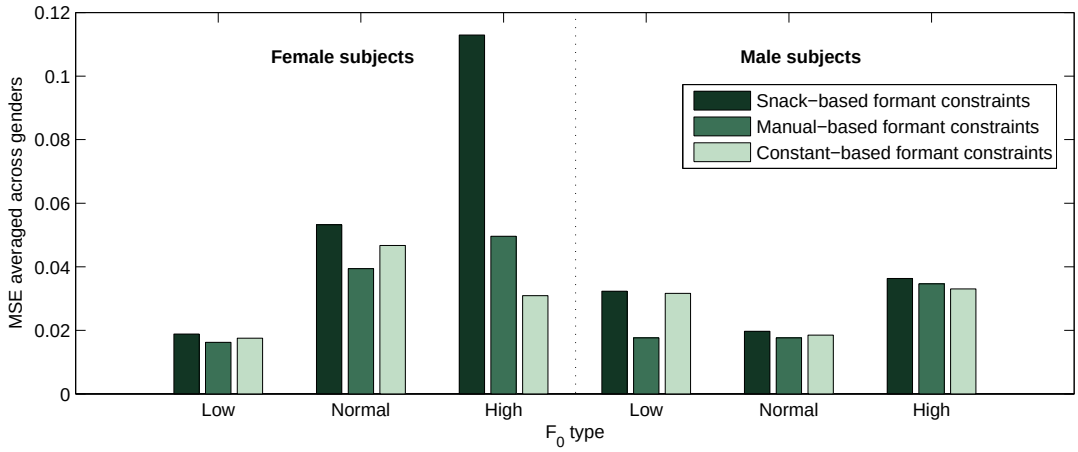



**Figure 3.6:** MSEs averaged across all phonations for each gender in terms of the $F_0$ type (low, normal and high) and type of formant constraint (Snack, manual and constant).

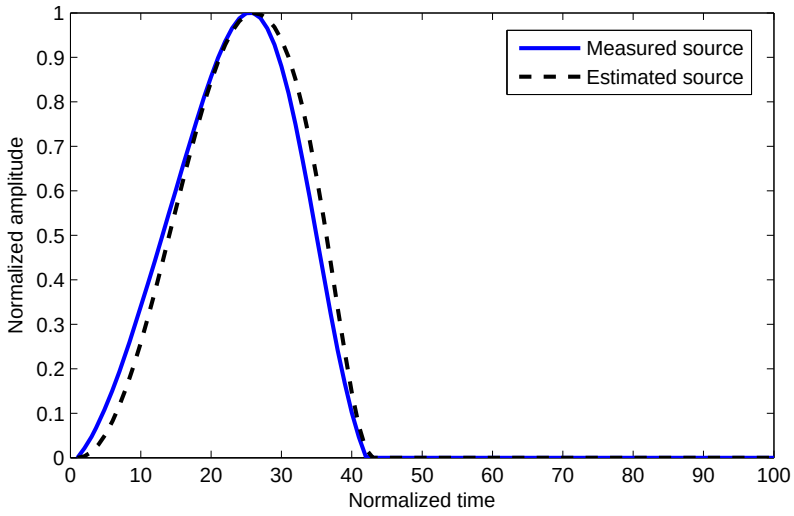



**Figure 3.7:** Phonation with the lowest source estimation error (MSE = 0.0018). The measured source waveform was taken from the high $F_0$, pressed phonation of subject FM1. The estimated source waveform (dashed) was from the manual-based formant constraints method.

to the normal and breathy phonations, although this was not seen for the female subjects. The manual- and constant-based formant constraints had similar error values for both genders.

Figure 3.6 shows the MSE values grouped in terms of the $F_0$ type for each gender. Again, it can be seen that, for both genders, the high $F_0$ phonations had the higher errors values, with the Snack-based formant constraints resulting in the highest error values within these phonations.

Figure 3.7 shows an example of a phonation with a low source estimation error (MSE = 0.0018). The measured source waveform was taken from the high $F_0$, pressed phonation of subject FM1 and the estimated source waveform (shown in the dashed line) was from the manual-based formant constraint method. Visual inspections of other estimated source waveforms showed that for the pressed

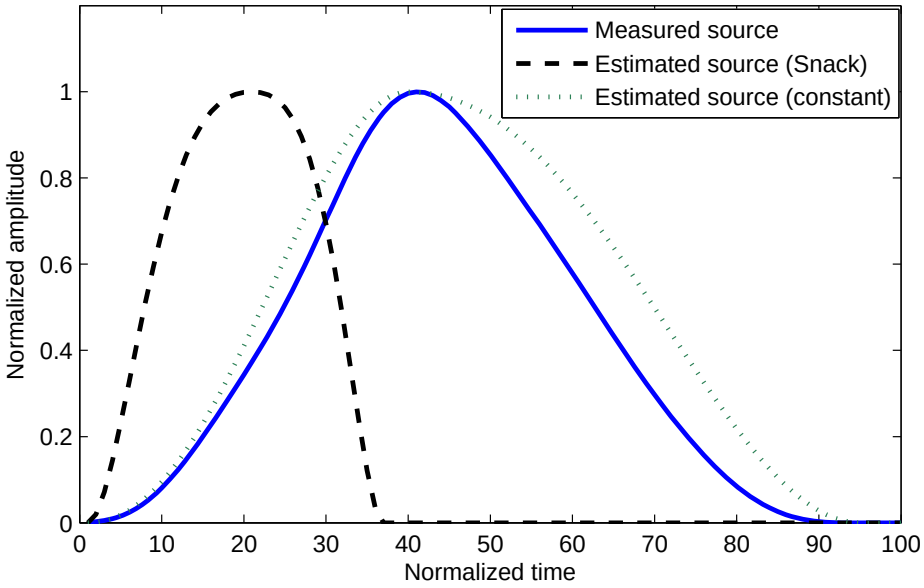



**Figure 3.8:** Phonation with the highest source estimation error. The measured source waveform was taken from the high $F_0$, breathy phonation of subject FM3 with the DC-offset removed. The dashed line shows the estimated waveform using Snack-based formant constraints (MSE = 0.2995) and the dotted line shows the estimated waveform using constant-based formant constraints (MSE = 0.0116).

cases, when the estimated $OQ$ was close to the measured $OQ$, the estimated source always provided a better fit than for other voice qualities. This may be because in most pressed phonations, the $OQ$ is smaller than normal and there is little time to change the vocal fold configurations to produce source pulse shapes with non-symmetrical attributes.

In contrast, Figure 3.8 shows an example of a phonation with a high source estimation error. The phonation was from subject FM3's high $F_0$, breathy utterance. The dashed line shows the estimated waveform using Snack-based formant constraints, with MSE = 0.2995, and the dotted line shows the estimated waveform using constant-based formant constraints, with MSE = 0.0116. This example illustrates a problem which can occur with incorrect formant constraints, where the true formant frequency lies outside the range specified by the constraints. In this case, the formant frequency constraints were too low, and the first formant was able to simulate the low frequency source information. When the higher constant-based formant constraints were used, a more accurate source was estimated.

## 3.5 Summary

In this chapter, an analysis-by-synthesis technique in the frequency domain is described which utilizes a codebook-search to effectively inverse-filter speech signals with the voice source signal. While the results are promising with respect to the source parameters, $OQ$, and to a lesser extent, $S_{op}$, the estimation errors with the remaining parameters, $\alpha$ and $S_{cp}$, suggest that not all parameters affect the harmonic magnitudes of the speech spectrum. Although this source estimation algorithm does not require precise formant frequency values, the analysis of the MSEs show that the formant frequency constraints need to contain the actual

formant frequencies in order to produce reasonable source estimates.

# CHAPTER 4

# Acoustic Correlates of Voice Quality

Voice quality is the component of speech which characterizes a speaker's voice, encompassing the effects of age, gender, prosody, and speaking styles. As shown in Chapter 3, precise voice source information extraction from the speech signal is not a solved problem. Instead of explicitly finding the voice source signal, we can use cues and measures which are correlated with certain aspects of the voice source. In this chapter, a new Matlab program named `VoiceSauce` is introduced which simplifies the calculation of these measures. This program is then used in three applications: voice source analysis with respect to voice quality correlates, gender identification, and prosody analysis.

## 4.1 Acoustic measures related to voice quality and to the voice source

Some of the more commonly used acoustic measures were reviewer in Section 1.4.1 and are listed in Table 4.1. In addition to these measures, the uncorrected version of the harmonic measures are also sometimes used; e.g. $H_1 - H_2$ instead of $H_1^* - H_2^*$. These are useful for data consisting of the same vowels.

$F_0$ is, by definition, a correlate of pitch accents and boundary tones. Pitch accents and boundary tones are deviations of the $F_0$ contour from a speaker's norm; this can be either a drop in $F_0$, as would happen in the case of a low pitch

accent or low boundary tone, or a rise in $F_0$, as in a high pitch accent or high boundary tone.

The measure $H_1^* - H_2^*$ and its uncorrected version, $H_1 - H_2$, has often been thought of as a correlate of open quotient ($OQ$), the proportion of time the vocal folds are open during a cycle of phonation. Inverse-filtered airflow and EGG measurements were used in [HHP95] to show a moderate correlation between estimated $H_1 - H_2$ and $OQ$ values. Since pressed and breathy voices are generally thought to be characterized by small and large $OQ$ values respectively ([Huf87, Fis67, SL90, Esp03]), $H_1^* - H_2^*$ has also been used as a correlate of breathiness. This relationship was partly supported by perceptual studies ([HCE94, KK90, Esp06]) which showed that listeners were more likely to judge phonations with a larger first harmonic magnitude as being breathy. However, in [HdD01], analysis of the LF model parameters showed that $H_1^* - H_2^*$ was dependent not only on $OQ$, but also on the asymmetry coefficient. In another study ([KGI08]), it was shown that the $OQ$ values obtained from high-speed imaging were only weakly correlated with $H_1^* - H_2^*$ values.

In [Han97], analysis of circuit models suggested that the spectral tilt measures $H_1^* - A_1$ and $H_1^* - A_3^*$ may be correlated with the speed of closure of the vocal folds. It was speculated that lower spectral tilt values should correspond to more abrupt glottal closures and higher values may be an indication of non-simultaneous closure. Fiberscopy of a small subset of speakers confirmed this to be the case, although in that study, the fiberscopic images were not collected simultaneously with the acoustic data. The measures $H_2^* - H_4^*$ and $H_1^* - A_2^*$ are also thought to be correlated with the source spectral tilt, at the mid-frequency range ([KGB07]). However, these correlations have not been verified experimentally.

*Energy* is, by definition, related to loudness and voice intensity, and may be

correlated with vocal effort. Other studies have shown that energy can be a good predictor of pitch accents and intonational boundaries ([CHC05, RH06, Sli07]).

The measures *CPP* and *HNR* have been perceptually related to modality and breathiness ([HCE94, Kro93, Esp06]). However, it is not clear how these measures are related physiologically to the voice source or how they affect the voice source model parameters.

**Table 4.1:** List of acoustic measures though to be related to the voice source and/or voice quality

| **Measure** | **Relation to voice quality or voice source** |
|---|---|
| $F_0$ | Correlated with pitch accents, boundaries and stress. |
| $H_1^* - H_2^*$ | Thought to be correlated with breathiness and open quotient ($OQ$, the proportion of time the vocal folds are open during phonation). |
| $H_1^* - A_3^*$ | Thought to be correlated with source spectral tilt and hence, the rate of closing of the vocal folds. |
| *Energy* | Related to loudness and voice intensity. |
| $H_2^* - H_4^*$, $H_1^* - A_1^*$, $H_1^* - A_2^*$ | Thought to be correlated with source mid-frequency tilt. |
| *CPP* | Cepstral peak prominence: thought to be correlated with modality vs. breathiness. |
| *HNR* | Harmonic-to-noise ratio: detects the amount of noise (turbulence or otherwise) in the speech signal. |

## 4.2 `VoiceSauce` - a program for voice analysis

`VoiceSauce` is a customizable application, implemented in Matlab, which provides automated acoustic measures over time from audio recordings. The measures currently computed are: $F_0$, $H_1(^*)$, $H_2(^*)$, $H_4(^*)$, $A_1(^*)$, $A_2(^*)$, $A_3(^*)$, $H_1(^*) - H_2(^*)$, $H_2(^*) - H_4(^*)$, $H_1(^*) - A_1(^*)$, $H_1(^*) - A_2(^*)$, $H_1(^*) - A_3(^*)$, *Energy*, *CPP*, *HNR*, formant frequencies $F_1$–$F_4$, and formant bandwidths $B_1$–$B_4$. `VoiceSauce` takes as input a folder of `.wav` files, and for each input `.wav` file produces a Matlab file with values every frame for all measures. It can operate on the whole input file, or over segments delimited by a Praat ([BW10]) textgrid file. `VoiceSauce` then takes these Matlab outputs, optionally along with electroglottographic (EGG) measurements from PCQuirerX, and provides condensed outputs in text format; alternatively it can write the Matlab outputs to the format used by the Emu Speech Database system [CH96]. `VoiceSauce` and its documentation can be obtained freely at: `http://www.ee.ucla.edu/~spapl/voicesauce`

### 4.2.1 $F_0$ and formant calculations

$F_0$ values in `VoiceSauce` can be calculated with the STRAIGHT algorithm [KCP98], the Snack sound toolkit [Sj04], or with an external application. The parameters, which are user-configurable in the settings dialog, are the window length, frame shift, maximum $F_0$ value, and minimum $F_0$ value.

Formant frequencies and bandwidths can also be calculated with the Snack sound toolkit or with an external application. The window length, frame shift and pre-emphasis factor are the user-configurable parameters.

### 4.2.2 Harmonic magnitudes and spectral amplitude calculations and corrections

The harmonic magnitudes, $H_1$, $H_2$ and $H_4$, are calculated by using the $F_0$ information to find the spectrum maximum near the frequencies $F_0$, $2F_0$ and $4F_0$. A variable window length corresponding to 3 pitch periods, as determined by the $F_0$ value at a particular point, is used to calculate the spectrum. The data are first multiplied by a Hamming window to reduce the effects of inaccurate $F_0$ estimates. Because only 3 harmonic magnitudes are required in each frame of data, optimization on a section of the spectrum is used instead of calculating the entire spectrum for each frame. The optimization seeks to find the maximum spectrum value around a particular frequency and is shown in Eq. 4.1 for the first harmonic, $H_1$.

$$H_1 = \max_{0.9f_0 \leq f \leq 1.1f_0} 20 \log \left( \left| \sum_{n=0}^{N} s(n) e^{-2\pi jnf/F_s} \right| \right) \tag{4.1}$$

where $s(n)$ are the data within a frame with window length $N$ and $f_0$ is the corresponding pitch frequency for that particular frame. The search range is $\pm 10\%$ of the given pitch frequency.

The spectral amplitudes at the formant frequencies, $A_1$, $A_2$ and $A_3$ were calculated in a similar way to the harmonic magnitudes, but used $F_1$, $F_2$ and $F_3$ values instead of $F_0$ values in the spectral maximum search. In this way, the amplitudes found will correspond to the largest harmonic amplitudes near the formant frequencies.

Correction for the effects of the vocal tract are performed using the following formula from [IA04]:

$$H^* = H(\omega) - \sum_{i=1}^{N} 10 \log_{10} \frac{\left(1 - 2r_i \cos(\omega_i) + r_i^2\right)^2}{\left(1 - 2r_i \cos(\omega + \omega_i) + r_i^2\right)\left(1 - 2r_i \cos(\omega - \omega_i) + r_i^2\right)}$$

where $H(\omega)$ is the magnitude of the actual signal spectrum (in dB) at frequency $\omega$, $N$ is the number of formants, $r_i = e^{-\pi B_i/F_s}$ and $\omega_i = 2\pi F_i/F_s$. $F_s$ is the sampling frequency, and $F_i$ and $B_i$ are the formant frequency and bandwidth for the $i$-th formant, respectively. Since the bandwidths from Snack have relatively large variances, the bandwidths are derived from the formant-to-bandwidth mapping formula in [HM95]:

$$B_i = S \cdot (k + (x_1 \cdot F_i) + (x_2 \cdot F_i^2) + (x_3 \cdot F_i^3) + (x_4 \cdot F_i^4) + (x_5 \cdot F_i^5))$$

where $F_i$ is the formant frequency and $k = 165.327516$, $x_1 = -6.73636734 \times 10^{-1}$, $x_2 = 1.80874446 \times 10^{-3}$, $x_3 = -4.52201682 \times 10^{-6}$, $x_4 = 7.49514000 \times 10^{-9}$ and $x_5 = -4.70219241 \times 10^{-12}$ for $F_i < 500$ Hz, and $k = 15.8146139$, $x_1 = 8.10159009 \times 10^{-2}$, $x_2 = -9.79728215 \times 10^{-5}$, $x_3 = 5.28725064 \times 10^{-8}$, $x_4 = -1.07099364 \times 10^{-11}$ and $x_5 = 7.91528509 \times 10^{-16}$ for $F_i \geq 500$ Hz. The scaling factor $S$ is defined as $S = 1 + 0.25\left(\frac{F_0 - 132}{88}\right)$ where $F_0$ is the pitch frequency. If selected, the harmonic magnitudes and spectral amplitudes are all corrected for the effects of formants $F_1$ and $F_2$. Measure $A_3^*$ is further corrected for the effects of formant $F_3$.

### 4.2.3 Energy calculation

For an input signal $s(t)$, the traditional energy measure is $s^2(t)$. However, for a fixed length analysis window, this measure would be correlated with $F_0$; i.e. the higher the $F_0$, the more pulses in the signal, and hence, the larger the energy value. To reduce this correlation, energy calculations in `VoiceSauce` are normalized for the effects of $F_0$. This is done by using a variable window length of 5 pitch periods, as determined by the $F_0$ value at the particular time instance.

### 4.2.4 *CPP* and *HNR* calculation

Cepstral peak prominence calculations are based on the algorithm described in [HCE94], while the harmonic to noise ratio measures are derived from [Kro93]. Similar to the *energy* measure, a variable window of length equal to 5 pitch periods is used for the calculations. After multiplying the data with a Hamming window, the data are then transformed into the real cepstral domain. The *CPP* is found by performing a maximum search around the quefrency of the pitch period ($1/F_0$). This peak is normalized to the linear regression line which is calculated between 1 ms and the maximum quefrency. The *HNR* measurements are found by liftering the pitch component of the cepstrum and comparing the energy of the harmonics with the noise floor. In `VoiceSauce`, the *HNR* is calculated for the frequency ranges 0–0.5 kHz, 0–1.5 kHz and 0–2.5 kHz and denoted by *HNR05*, *HNR15* and *HNR25*, respectively.

## 4.3 Application I: Voice quality analysis with respect to acoustic measures

Acoustic measures provide a way to analyze the voice source without explicitly estimating the voice source signal. These measures are typically found in the spectral domain and can be used in voice quality applications such as prosody analysis and detection, speaker identification, as well as various medical applications. The measures reviewed in Section 1.4.1 are the most commonly used measures in studies of the voice source and voice quality. However, with the exception of $H_1^* - H_2^*$, no other acoustic measure has been strongly associated with the physiological attributes of the voice source over a wide range of voice qualities.

The linkages between the voice source signal, the perceived voice quality and measures related to the voice source are, to date, not well understood. This general disconnect can be mainly attributed to the lack of data from the direct observations of the voice source, which makes it difficult to form the connections between the voice source and their acoustic correlates. In this section, high-speed imaging of the vocal folds along with simultaneous audio recordings are used to test the relationships between three voice qualities (pressed, normal and breathy), the voice source shape, and the acoustic measures related to the voice source.

### 4.3.1 Data

The data used in this work are the same as those described in Section 2.2 and summarized here. Synchronized audio recordings and high-speed imaging of the vocal folds were collected from six subjects (3 males, denoted by M1–3, and 3 females, denoted by FM1–FM3). The subjects were asked pronounce the vowel /i/ while varying their $F_0$ (low, normal and high) and voice quality (pressed, normal and breathy) in a quasi-orthogonal way; this resulted in 9 phonations for each subject. One male subject, M1, was unable to produce a low $F_0$ phonation for any voice quality and one female subject, FM2, was unable to produce a low $F_0$, normal phonation. The total number of phonations was 50.

### 4.3.2 Methods

The voice source signal, as represented by the glottal area waveform, was obtained through the process described in Section 2.2.3. These waveforms are then fitted to the proposed new source model specified in Section 2.3, allowing the voice source to be compactly represented with four parameters: $OQ$ (open quotient), $\alpha$ (asymmetry coefficient), $S_{op}$ (speed of opening phase), and $S_{cp}$ (speed of closing

**Table 4.2:** Occurrences of glottal gaps in terms of speaker, $F_0$ type (low, normal and high) and voice quality (pressed, normal and breathy). '–' denotes an entry where no speaker produced a glottal gap.

| $F_0$ **type** | **Voice quality** | | |
|---|---|---|---|
| | Pressed | Normal | Breathy |
| Low | – | FM1, FM2, FM3 | FM1, FM2, FM3, M2 |
| Normal | FM3 | FM2 | FM1, FM2, FM3, M1, M2, M3 |
| High | M1 | M3 | FM1, FM2, FM3, M1, M2, M3 |

phase). For this study, the model parameter $OQ$ was defined as the time from the first opening instant to the onset of maximum closure, divided by the fundamental period; this definition ignores the glottal gaps which were recorded separately.

The phonations with incomplete glottal closures, as represented by a DC-offset in the glottal area waveforms (see Appendix A), were manually marked. The occurrences of the glottal gaps are shown in Table 4.2. On average, it can be seen that the female subjects had more phonations with glottal gaps than the male subjects and the breathy phonations result in more incomplete glottal closures than the other voice qualities.

Before calculating the acoustic measures, the audio recordings were all normalized to the maximum amplitude to ensure the energy measures were not affected by the recording level. The `VoiceSauce` application (Section 4.2) was then used to calculate the following measures: $H_1-H_2$, $H_2-H_4$, $H_1-A_1$, $H_1-A_2$, $H_1-A_3$, *CPP*, *HNR05*, *HNR15*, *HNR25* and *energy*. Since the phonations within each subject were all approximately the same vowel, only the uncorrected versions of the harmonic magnitudes and spectral amplitudes were used in the analysis. Each measure was calculated with the default frame shift size of 1 ms and averaged

over the whole phonation.

Statistical analyses were performed with the software package SPSS (v16.0). For the two-way analysis of variances (ANOVA) tests, fixed factors included the speaker plus one other factor from voice quality, glottal gap and $F_0$ type. Tests where the null hypothesis had a probability of $p < 0.001$ were considered to be statistically significant.

### 4.3.3 Results

#### 4.3.3.1 Voice quality effects

Statistical analysis was performed on all of the phonations, using voice quality (pressed, normal and breathy) as one of the fixed factors. The voice source model parameters and acoustic measures which showed statistical significance are listed in Table 4.3; values listed are the $F$ value[1] with its associated degrees of freedom, $\eta^2$ (measure of effect size), and the parameter/measure means and standard deviations. Not surprisingly, $OQ$ was shown to be lowest for the pressed phonations and highest for the breathy phonations. This is in agreement with existing studies ([Huf87, Fis67, SL90]) which suggested this effect. Figure 4.1 shows the mean $OQ$ values for each subject averaged over the pressed, normal and breathy phonations. With the exception of M1, the other subjects all had the same trend for the $OQ$ value: pressed $<$ normal $<$ breathy. An inspection of the glottal area waveform for subject M1 (see Appendix A) shows that the pressed, high $F_0$ phonation has a DC-offset. This may be because the high $F_0$ produced does not give the vocal folds enough time to return to the closed phase,

[1]The $F$ value is defined as the ratio of the model mean square to the error mean square and the partial $\eta^2$ value is calculated as $SS_{effect}/(SS_{effect} + SS_{error})$, where $SS_{effect}$ is the sum of squares of the effect and $SS_{error}$ is the sum of squares of the error.

**Table 4.3:** Voice source model parameters and acoustic measures which were affected by voice quality in a statistically significant way. Values shown are the $F$ value (ratio of the model mean square to the error mean square), $\eta^2$ (measure of the effect size), and the parameter/measure means and standard deviations (in parentheses) for the three voice qualities.

| | $F$ **and** $\eta^2$ | | **Mean (s.d.) of parameter/measure** | | |
|---|---|---|---|---|---|
| **Parameter** | $F(1,2)$ | $\eta^2$ | Pressed | Normal | Breathy |
| $OQ$ | 28.568 | .641 | .650(.128) | .804(.128) | .935(.062) |
| $\alpha$ | 31.576 | .664 | .507(.057) | .486(.088) | .386(.041) |
| **Measure** | | | | | |
| *CPP* | 27.935 | .636 | 25.081(3.290) | 23.991(2.396) | 18.042(2.744) |
| *HNR05* | 9.331 | .368 | 15.414(10.559) | 13.502(7.817) | 3.442(6.522) |
| *HNR15* | 9.633 | .376 | 24.591(10.537) | 23.435(5.762) | 13.325(6.868) |
| *HNR25* | 10.340 | .393 | 27.406(10.099) | 26.311(5.841) | 16.401(6.380) |
| $H_1 - A_2$ | 8.871 | .357 | 13.502(7.054) | 17.266(8.593) | 23.025(6.503) |
| $H_1 - A_3$ | 18.099 | .531 | 20.499(6.158) | 24.347(6.724) | 29.796(6.322) |
| $H_1 - H_2$ | 16.641 | .510 | -0.215(6.791) | 1.670(6.213) | 11.188(4.583) |

resulting in a larger $OQ$ than normal for the pressed phonation.

A somewhat unexpected result was seen for the model parameter $\alpha$, although a post-hoc analysis showed that the main effect was due to the pressed/normal vs. breathy voice qualities. The results showed that, on average, the pressed and normal phonations were more symmetrical (equal durations for opening and closing phases) than the breathy phonations which were skewed towards a shorter opening phase. This is demonstrated in Figure 4.2 using the mean $OQ$ and $\alpha$ values in Table 4.3 with $S_{op}$ and $S_{cp}$ set to 0.5. This result is surprising because

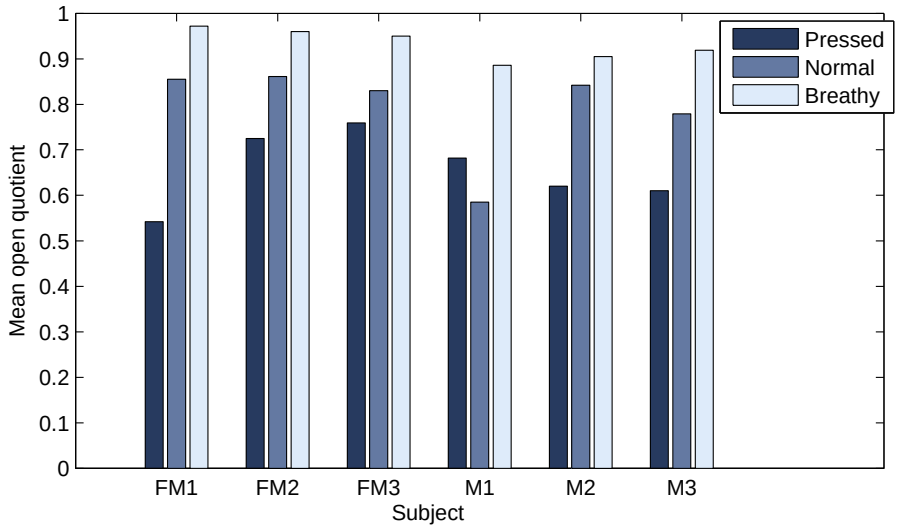



**Figure 4.1:** Mean $OQ$ values for each speaker averaged over the pressed, normal and breathy phonations.

the duration of the opening phase has conventionally been thought to be always longer than the duration of the closing phase, due to the effort require to separate the vocal folds, and also because this is what has been seen in EGG and airflow signals. Individual subject analysis showed that all subjects had the lowest $\alpha$ values for the breathy phonations.

The statistical analysis on the acoustic measures showed that most of the measures (*CPP*, *HNR*, and $H_1 - H_2$) thought to be related to breathiness were statistically significant. However, post-hoc analysis on these measures revealed that there were few differences between the pressed and normal phonations, with the statistical significance coming mainly from the pressed/normal vs. breathy phonations. For the *CPP* measure, the mean values were higher for the pressed and normal phonations than for the breathy phonation. This was as predicted in [HCE94], and could be attributed to the rising of the noise floor in the speech spectrum for breathy phonations. Similarly, the *HNR* (*HNR05*, *HNR15* and *HNR25*) measures were much lower for the breathy phonations due to increased

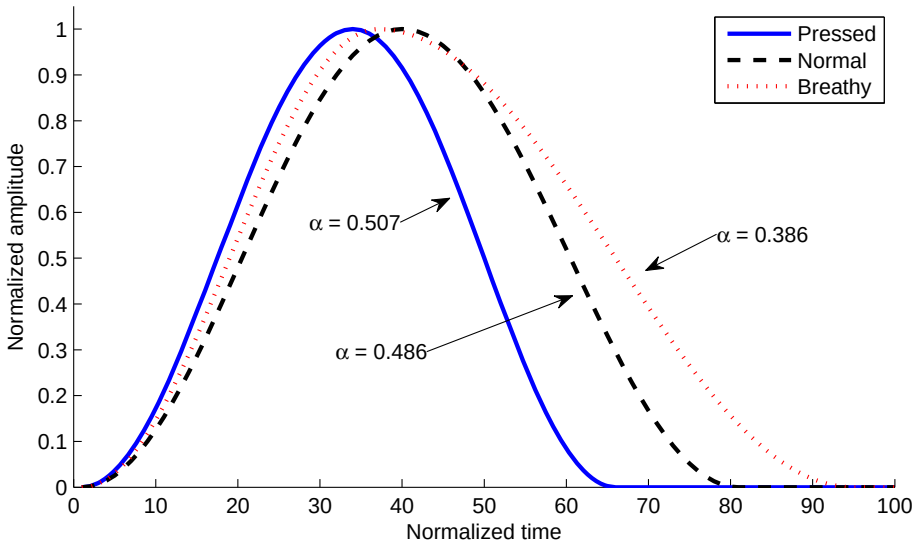



**Figure 4.2:** Examples of voice source shapes for the mean $OQ$ and $\alpha$ values listed in Table 4.3; $S_{op}$ and $S_{cp}$ were both set to a value of 0.5.

noise in the spectrum. Interestingly, the $H_1 - H_2$ measure had similar means for the pressed and normal phonations, but a significantly larger value for the breathy phonation. This is slightly different from the trends for the $OQ$ parameter which had progressively increasing values from the pressed to normal to breathy voice qualities.

On average, the spectral tilt measures $H_1 - A_2$ and $H_1 - A_3$ were smallest for the pressed phonation and largest for the breathy phonation. These results confirm the hypothesis in [Han97] that voice sources with more abrupt glottal closures may lead to more high frequency components in the speech spectrum. A similar, but indirect, result in [SV96b] found that tenser, stressed phonations had more high frequencies than lax phonations.

Individual subject analysis for the acoustic measures found that all six subjects had similar trends for the measures.

#### 4.3.3.2 Glottal gap effects

The results in Section 4.3.3.1 showed that for the parameter $\alpha$ and the voice source related measures, there were few differences separating the pressed and normal phonations. However, the breathy phonations were seen to have significantly different values from either the pressed/normal phonations. A possible cause for this effect could be due to the existence of glottal gaps for the breathy phonations. 16 out of the total 17 breathy phonations had glottal gaps, with the exception being the breathy, low $F_0$ phonation for subject M3.

Table 4.4 lists the model parameters and acoustic measures which were statistically significant in the ANOVA analysis, with the presence/absence of the glottal gap as the other fixed factor, along with the subject. Given that glottal gaps usually occurred with the breathy phonations, it was not surprising to see the $OQ$ parameter being associated with the glottal gap effect. Similarly, it was shown previously that the $\alpha$ parameter had the lowest mean value for the breathy phonations, hence the statistical significance with the glottal gap factor. While it can be seen from these results that $OQ$ is dependent on both the type of voice quality (pressed, normal or breathy) and the existence/absence of the glottal gap, it is not clear as to how or which factor is predominantly affecting $\alpha$, or if both factors are affecting it, similar to $OQ$. Analysis of the phonations which contained a glottal gap and were not of a breathy voice quality showed that the $\alpha$ values for these phonations were not necessarily the lowest for their corresponding voice quality group. However, for all subjects, the breathy phonation had the lowest $\alpha$ values when averaged across each subjects' $F_0$ type. From these results, it would be reasonable to hypothesize that it is the breathy phonations which affect the $\alpha$ values, but more data would be needed to confirm this.

Interestingly, parameter $S_{op}$, which did not show a statistically significant

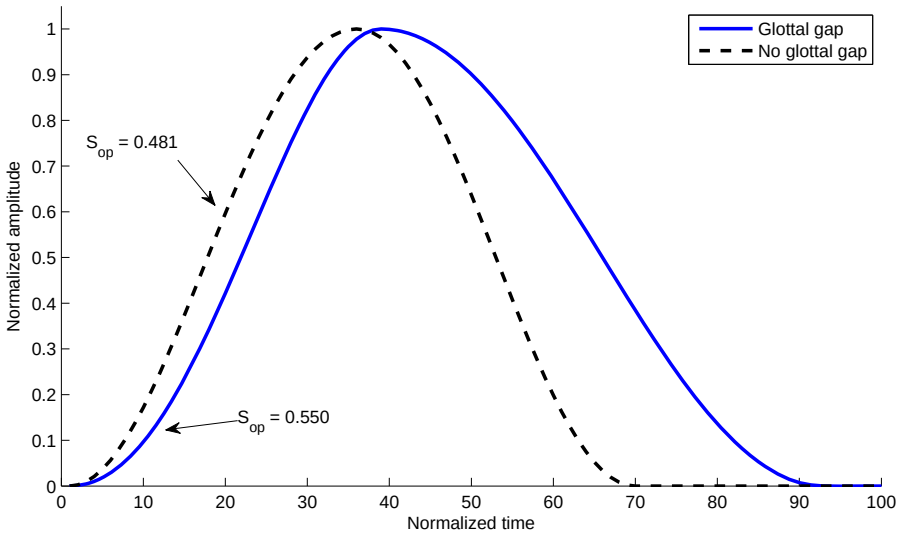



**Figure 4.3:** Examples of voice source shapes for the mean $OQ$, $\alpha$ and $S_{op}$ values listed in Table 4.4; $S_{cp}$ was set to a value of 0.5.

effect of voice quality, showed a statistical significant effect of the glottal gap. The larger mean value for the presence of the glottal gap translates to a slower initial rise during the opening phase. This is shown in Figure 4.3 using the mean values in Table 4.4 for parameters $OQ$, $\alpha$ and $S_{op}$; $S_{cp}$ was set to a value of 0.5. A possible explanation for the slower initial rise during the opening phase could be due to the smaller distance required to reach the maximum open position of the vocal folds. Without the glottal gap, the distance from the closed position to the maximum open position is much greater, hence requiring a faster initial rise during the opening phase. Another way of interpreting these results could be that the opening phase is dictated by a constant "curve", and the glottal gap simply moves the starting point up this curve. This interpretation would also result in a lower $S_{op}$ value when using the full range of the curve and a high value when starting near the middle of this curve.

With the exception of the three *HNR* measures, the acoustic measures which

**Table 4.4:** Voice source model parameters and acoustic measures which were statistically significant to the effects of the glottal gap. Values shown are the $F$ value, $\eta^2$, and the parameter/measure means and standard deviations (in parentheses) for the phonations with glottal gaps and without glottal gaps.

| | **$F$ and $\eta^2$** | | **Mean (s.d.) of parameter/measure** | |
|---|---|---|---|---|
| **Parameter** | $F(1,1)$ | $\eta^2$ | Glottal gap | No glottal gap |
| $OQ$ | 47.480 | .555 | .922(.068) | .694(.139) |
| $\alpha$ | 38.690 | .504 | .413(.066) | .498(.077) |
| $S_{op}$ | 15.414 | .289 | .550(.066) | .481(.074) |
| **Measure** | | | | |
| $CPP$ | 24.677 | .394 | 19.631(3.515) | 24.605(3.258) |
| $H_1 - A_2$ | 27.039 | .416 | 22.070(6.812) | 14.569(7.958) |
| $H_1 - A_3$ | 44.184 | .538 | 29.451(6.127) | 21.150(6.116) |
| $H_1 - H_2$ | 29.089 | .434 | 9.371(5.937) | -0.014(6.283) |

were statistically significant to the voice quality factor were also statistically significant to the presence/absence of the glottal gap. This is not surprising given that the same measures appear to be predominantly affected by the breathy voice quality which contains most of the phonations with glottal gaps. The mean values for the measures *HNR05*, *HNR15* and *HNR25* were also lower, inferring more noise, for the presence of the glottal gap, but these were not statistically significant. This suggests that noise may be more prevalent in breathy phonations as opposed to phonations with incomplete glottal closures, which may or may not be breathy. A related study ([KK90]) found that, during perceptual experiments, listeners were more likely to rate a phonation as breathy if an increase in $H_1 - H_2$ was accompanied by noise; increases in $H_1 - H_2$ alone were sometimes rated as having a nasalized voice quality. Similarly, in [Kha09], it was found that $H_1^* - H_2^*$ only worked for some speakers in separating breathy vs. modal vowels in Gujariti.

In [Han97], it was suggested that speakers with high $H_1^* - A_1$ and $H_1^* - A_3^*$ values may have a posterior opening in the vocal folds. This hypothesis is supported here by the related measure, $H_1 - A_3$, which has a high mean value for the glottal gap case. Although the mean values for $H_1 - A_1$ also showed the same trend, the effect was not statistically significant.

#### 4.3.3.3 $F_0$ type effects

No voice source model parameters or acoustic measures were affected by the three $F_0$ types (low, normal and high) in a statistically significant manner. In terms of model parameters, this is not so surprising as shown by the recorded phonations, the three different voice qualities could be produced at three different levels of $F_0$ for most of the subjects. However, previous studies ([ISE06, ISA07]) utilizing more natural speech, have shown that there are correlations between $F_0$ and

certain acoustic measures such as $H_1^* - H_2^*$ and energy. In this work, the data used were more of a static nature, and this may have reduced the effects of $F_0$ on these measures.

#### 4.3.3.4 Correlations between model parameters and acoustic measures

Table 4.5 lists the correlations ($r$) between voice source model parameters ($OQ$, $\alpha$, $S_{op}$ and $S_{cp}$) and the acoustic measures (*CPP*, *HNR*, $H_1-A_1$, $H_1-A_2$, $H_1-A_3$ and $H_1-H_2$). Measures $H_2-H_4$ and *energy* did not show any strong correlations with model parameters. Parameter $S_{cp}$ also did not show any meaningful correlation but is listed for comparison with $S_{op}$.

It can be seen that the parameter $OQ$ is moderately correlated with the parameters $\alpha$ and $S_{op}$, and also with the measures *CPP*, $H_1 - A_1$, $H_1 - A_2$, $H_1 - A_3$ and $H_1 - H_2$. The correlations with $\alpha$ and $S_{op}$ were not surprising given that $\alpha$ appeared to be affected by voice quality and $S_{op}$ by the presence/absence of the glottal gap, both effects which were correlated with $OQ$. The negative correlation with *CPP* is most likely attributable to the breathy voice quality; since breathy phonations were seen to induce larger $OQ$ values and also more spectral noise, hence resulting in a smaller *CPP* value. Correlations with the spectral tilt measures $H_1 - A_1$, $H_1 - A_2$ and $H_1 - A_3$ could be explained using the reasoning from [Han97]; that is, when the glottal closures become less abrupt, as in the case when $OQ$ increases, the high frequency components are generally reduced. The moderate correlation with $H_1 - H_2$ was as predicted by [HHP95], although the correlation here was not quite as strong as that study ($r = 0.6563$ vs. $r = 0.6928$). However, as shown by the mean values in Table 4.3, the mean $H_1 - H_2$ values did not increase linearly for the three types of voice qualities as

**Table 4.5:** Correlations between voice source model parameters and acoustic measures. Values are the correlation coefficients ($r$); correlations with $r > 0.4$ are in bold and were all statistically significant. Measures $H_2 - H_4$ and *Energy* did not show any meaningful correlations with any voice source parameters.

| Parameters/Measures | **Voice source model parameters** | | | |
|---|---|---|---|---|
| | $OQ$ | $\alpha$ | $S_{op}$ | $S_{cp}$ |
| $\alpha$ | **-0.5546** | – | – | – |
| $S_{op}$ | **0.5034** | -0.3306 | – | – |
| *CPP* | **-0.5445** | **0.5256** | -0.1617 | -0.1240 |
| *HNR05* | -0.3187 | **0.4112** | -0.0985 | -0.1339 |
| *HNR15* | -0.3536 | **0.4151** | -0.1814 | -0.0685 |
| *HNR25* | -0.3370 | **0.4606** | -0.1521 | -0.0798 |
| $H_1 - A_1$ | **0.4998** | -0.3053 | 0.2452 | -0.0734 |
| $H_1 - A_2$ | **0.4454** | -0.3170 | 0.0808 | 0.0854 |
| $H_1 - A_3$ | **0.5520** | -0.2250 | 0.0957 | 0.0522 |
| $H_1 - H_2$ | **0.6563** | **-0.4730** | 0.2641 | 0.1562 |

occurred with the parameter $OQ$. Furthermore, $H_1 - H_2$ also showed a slight correlation with the asymmetry coefficient, $\alpha$. This is similar to the findings in [HdD01], which used the LF model to theoretically show that $H_1^* - H_2^*$ was dependent on both $OQ$ and the asymmetry coefficient.

Apart from the measure $H_1 - H_2$, $\alpha$ was also correlated with *CPP*, and the three *HNR* measures. *CPP* was also moderately correlated with the parameter $OQ$ which was affected by the voice quality. Interestingly, the *HNR* measures were more strongly correlated with $\alpha$ than $OQ$, although the correlations are moderately weak for both parameters. This is not surprising since it was shown previously that both $\alpha$ and the *HNR* measures were thought to be predominantly affected by the breathy voice quality.

The lack of any meaningful correlations with the parameter $S_{cp}$ is surprising given that the parameter $S_{op}$ is moderately correlated with $OQ$; the correlation coefficient between $OQ$ and $S_{cp}$ is $r = 0.1825$ compared with $r = 0.5034$ for $S_{op}$. Since the tension of the laryngeal muscles is assumed to be constant during a cycle of phonation, this result requires further exploration.

### 4.3.4 Summary

In this work, direct measurements of the glottal area waveforms were used to examine the voice source model parameters and acoustic measures in relation to the effects of voice quality, glottal gaps and $F_0$. Using ANOVA tests, it was found that the model parameter $OQ$ and the spectral tilt measures $H_1 - A_2$ and $H_1 - A_3$ were affected by both voice quality and glottal gaps, while the parameter $\alpha$ was predominantly affected by voice quality, especially of the breathy type. This was also the case with many of the acoustic measures, such as *CPP* and the three *HNR* measures, indicating the presence of more spectral noise

for breathy phonations. Correlation analysis showed that the measure $H_1 - H_2$ was correlated with both the parameters $OQ$ and $\alpha$, which agrees with existing theoretical studies. However, the correlation between $OQ$ and $S_{op}$ and the lack of correlation between $OQ$ and $S_{cp}$ is puzzling and requires further research to unravel their relationship.

## 4.4 Application II: Automatic gender classification

Gender-based differences in human speech are due in part to physiological differences such as vocal fold thickness or vocal tract length, and differences in speaking style. Physiological properties of the glottis and the vocal tract change with age and gender. Since these changes are reflected in the speech signal, acoustic measures related to those properties can be helpful for age and gender classification. Assuming the linear source-filter model of speech production [Fan70], the contribution of acoustic measures to such classification can then be attributed to the voice source or the vocal tract. To our knowledge, with the exception of fundamental frequency ($F_0$), there has been no study that has examined the role of voice source related measures on age and/or gender classification.

It is well known that $F_0$ values for male talkers drop during adolescence due to a lengthening and thickening of their vocal folds. $F_0$ for adult males is typically around 120 Hz, while $F_0$ for adult females is around 200 Hz [PB52], similar to children.

It is also well known that, due to vocal tract length differences, adult males exhibit lower formant frequencies than adult females [PB52]. Interestingly, for preadolescent children, studies also found lower formant frequencies for boys compared to girls of ages 5-6 [WB71], 7-8 years [Ben80], and ages 5, 7, 9, and 11

years for Australian English [BP95]. These findings imply that, overall, boys have larger vocal tracts than girls. In [POA01], statistical analysis of children speech confirmed that formant frequencies ($F_1$, $F_2$, $F_3$), and not $F_0$, differentiate gender for children as young as 4 years of age, while formant frequencies plus $F_0$ differentiate gender after 12 years of age. These findings lead to the conclusion that for preadolescent children, vocal tract measures play a bigger role for gender classification than the voice source measure $F_0$. For adult speech, automatic gender classification has been presented in [WC91], which used linear predictive coding (LPC)-derived measures that represent the vocal tract.

In [LPN99], changes in magnitude and variability of, among other measures, $F_0$, formant frequencies, and spectral envelope are presented as a function of age for talkers from 5 to 50 years old. For $F_0$, the study showed a drop between ages 12 and 15 for males and a drop of $F_0$ variation for all talkers between ages 5 and 15. Formant frequencies ($F_1$, $F_2$, $F_3$) decreased between ages 10 and 15, where formant frequencies of male talkers decreased faster and reached much lower absolute values than those of female talkers. The study showed that children younger than age 10 displayed greater spectral variability than adults.

In [ISA07], age, sex, and vowel dependencies, were analyzed for talkers between the ages of 8 and 39 for the following three voice source related measures: $F_0$; $H_1^* - H_2^*$, and $H_1^* - A_3^*$. For male talkers, the results showed a drop of about 5 dB in $H_1^* - H_2^*$ around age 15 and a continuous decrease of $H_1^* - A_3^*$ between ages 8 and 39 by about 10 dB. For female talkers, the value of $H_1^* - H_2^*$ remained relatively unchanged between ages 8 and 39, whereas for $H_1^* - A_3^*$ a slight decrease by about 4 dB was shown. These developmental changes resulted in higher values of $F_0$, $H_1^* - H_2^*$, and $H_1^* - A_3^*$ for adult female talkers compared to adult male talkers [HC99].

In this section, acoustic measures from both the voice source and the vocal tract were used for automatic gender classification of 8 to 39 year old talkers. The vocal tract measures consist of formant frequencies and formant bandwidths, and the voice source measures used were $F_0$, $H_1^* - H_2^*$, and $H_1^* - A_3^*$. Training and testing was done using support vector machines (SVMs). The results were analyzed to see if voice source measures can improve automatic gender classification. Finally, the SVM classification results were compared with human perception classification tests, and also with classification results using conventional Mel-frequency cepstral coefficient (MFCC) features in combination with Gaussian mixture models (GMMs).

### 4.4.1 Speech data

Speech recordings from five age groups, ages 8–9, 10–11, 12–13, 14–15 and 16–39 were taken from the CID database [MLU96]. Each recording was of the form "I say uh, bVt again", where the target vowel 'V' was /ih/, /eh/, /ae/, /uw/ or /iy/. The word 'bead' was used to elicit the vowel /iy/ in this database. These utterances were spoken at the habitual speaking level and most talkers repeated the phrases twice. For the analysis here, only the manually segmented target vowels were used. The distribution of talkers (males/females) and number of utterances per age group is listed in Table 4.6. The total number of male/female talkers was 205/160 and the total number of utterances was 3880.

### 4.4.2 Methods

The acoustic measures used for gender classification were the first three formant frequencies ($F_1$, $F_2$, and $F_3$), the first two formant bandwidths ($B_1$ and $B_2$), and the measures related to the voice source $F_0$, $H_1^* - H_2^*$, and $H_1^* - A_3^*$. The mea-

**Table 4.6:** Distribution of gender and utterances for each age group.

| Age group | males/females | No. of utterances |
|---|---|---|
| 8-9 | 48/36 | 810 |
| 10-11 | 48/33 | 807 |
| 12-13 | 38/34 | 708 |
| 14-15 | 22/21 | 413 |
| 16-39 | 49/36 | 1142 |

sures were processed using `VoiceSauce` (Section 4.2) with the following settings: formant frequencies and bandwidth values were estimated with Snack (analysis window of 25 ms, frame shift of 1 ms and pre-emphasis factor of 0.96), and $F_0$ was estimated using the STRAIGHT algorithm. For each of the voice source measures, a first order Legendre polynomial was fitted to the raw values to obtain a measure of the mean and the slope (denoted by $\triangle$) across the duration of the vowel.

Classification was done using an SVM classifier with a Radial Basis Function kernel. In this study, the LIBSVM toolkit [CL01] was used to train and test on vectors containing different combinations of acoustic measures extracted from the five target vowels. For each classification experiment, 70% of the utterances, selected randomly, were used for training; the remaining utterances were used for testing. Five experiments were performed for each combination of acoustic measures and the average accuracy recorded.

For perception tests, four male subjects between ages 26 and 39 participated. They were each presented with 100 utterances of the target words and had to decide between male or female voice. The target words were manually segmented from the carrier phrase and were played back in random order using headphones.

The distribution of male and female utterances per age group are listed in Table 4.7. The same perception tests were also performed using just the segmented vowel part of the target word.

To compare the SVM results with more traditional methods, the first 12 MFCCs were extracted from the utterances and combined with the mean $F_0$ for each of the utterances to form a 13-dimension feature vector. Training was done with 2 GMMs each with 6 mixtures.

**Table 4.7:** Distribution of utterances used in perception experiments.

| Age group | No. of utterances male/female |
|---|---|
| 8-9 | 7/7 |
| 10-11 | 8/8 |
| 12-13 | 8/8 |
| 14-15 | 12/10 |
| 16-39 | 15/17 |

### 4.4.3 Results and discussion

For this section, the set of acoustic measures containing formant information ($F_1$, $F_2$, $F_3$, $B_1$, and $B_2$) will be denoted by FB.

#### 4.4.3.1 Results using $F_0$ and formants

As a first step, we analyzed the contribution to gender classification accuracy of only $F_0$, only FB, and $F_0$ plus FB (labeled by M0). These measures are the most widely used in gender and age classification. Figure 4.4 shows the classification accuracy for each age group using those measures. For ages 8 to 11 it can be

seen that formant information only (FB) performed slightly better than $F_0$. This is consistent with [POA01]. Gender classification accuracy for ages 8 to 13 was always below 65%, but between age groups 12–13 and 14–15, it increased to 85% for $F_0$ and to 68% for FB; these results can be attributed to the large drop of $F_0$ for males around ages 12 to 15 (about 105 Hz on average) [ISA07, LPN99] and to a decrease of formant frequencies for males relative to females [LPN99]. Since M0[2] overall yielded the best results, it was chosen as the baseline measure set for the comparison of the performance of voice source measures.

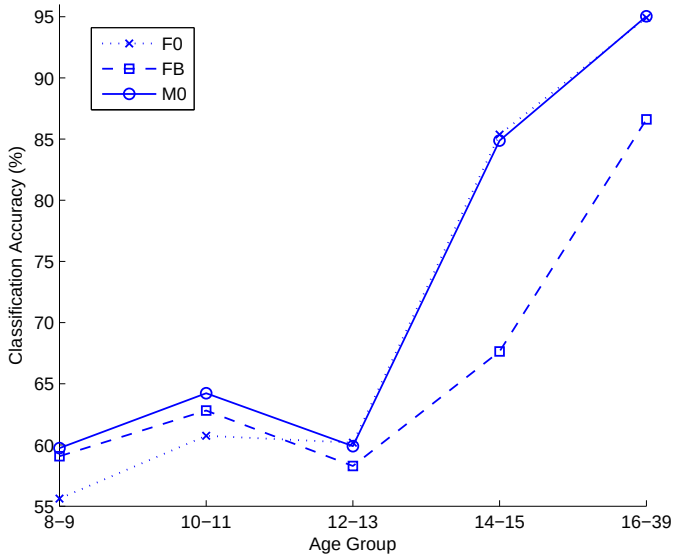



**Figure 4.4:** Gender classification accuracy for each age group using just $F_0$, just FB, and $F_0$ plus FB (M0).

#### 4.4.3.2 Results adding voice source measures

Figure 4.5 compares the changes in gender classification accuracies resulting from the addition of the various voice source measure sets (M1–M3) as listed in Ta-

[2]We also tested adding $\triangle F_0$ as a feature, but the classification accuracies were slightly lower when this measure was added.

**Table 4.8:** Measure sets (M0–M3) used in the gender classification tests. M0, in bold, is used as the baseline measure set.

| Set | Acoustic Measures | | | | | |
|---|---|---|---|---|---|---|
| | $F_0$ | FB | $H_1^* - H_2^*$ | $H_1^* - A_3^*$ | $\triangle F_0$ | $\triangle H_1^* - H_2^*$ |
| **M0** | ✓ | ✓ | | | | |
| M1 | ✓ | ✓ | ✓ | | | |
| M2 | ✓ | ✓ | ✓ | ✓ | | |
| M3 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

ble 4.8. The baseline measure set (M0) is shown as a solid line. Table 4.9 shows the values corresponding to this figure as well as results from MFCC/GMM classification tests. It can be seen that adding voice source measures plays a significant role only for age groups 10–11 and 12–13, where the absolute accuracy was improved by up to 9% using measure set M3. For age group 8–9, the accuracies were below 60% and the SVM seemed unable to model the classes for males and females satisfyingly. Although it was shown in [ISA07] that the source measures $H_1^* - H_2^*$ and $H_1^* - A_3^*$ are dependent on age and gender, the changes in classification accuracy for age groups 14–15 and 16–39 when using M1 or M2 were not significant. This could be attributed to the already large classification accuracy of the baseline (M0). Interestingly, while the classification accuracies for the voice source measure sets were similar to the MFCC/GMM results for age groups 8–9, 12–13 and 16–39, the voice source measure set performance for M2 is about 9% and 5% higher for age groups 10–11 and 14–15, respectively.

A closer look at the classification accuracy results for age group 12–13 is shown in Table 4.10, which shows the percentage correct classification of males and females. Compared to M0, the addition of the voice source measures assists in

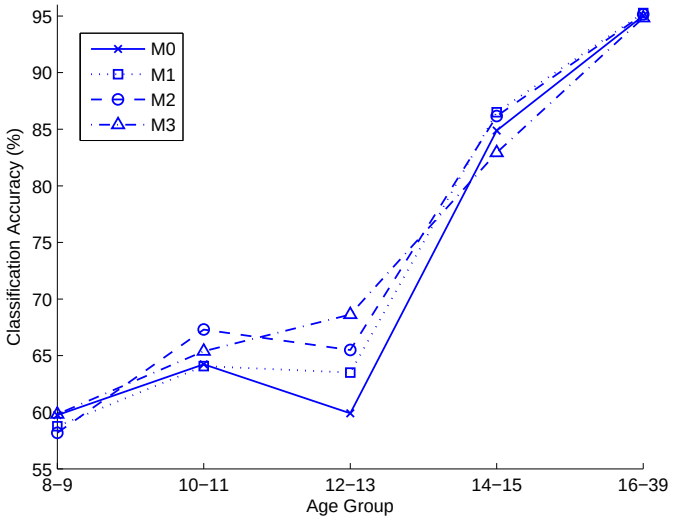



**Figure 4.5:** Gender classification accuracy for each age group using the measures sets M1, M2 and M3. M0 represents the baseline performance results. The corresponding values are listed in Table 4.9.

increasing the classification accuracy by about 7% for males and 9% for females when using M3. However, since the M2 measures are easier to calculate than those of M3, and M2 showed a classification accuracy improvement for all ages between 10 and 39, it is recommended to use M2 for gender classification. M2 will be used throughout the remainder of this results section.

#### 4.4.3.3 Comparison with perception results

Table 4.11 compares automatic classification results (denoted by AUT) with human perception results from this study (denoted by PER1) and from perception experiments in [POA01] (denoted by PER2). Note in [POA01], the target words were in a different context (hVd instead of bVt). These perception experiments were done using the target words. All values are gender recognition accuracies in percent. Dashes in the table represent unavailable data. AUT results were from

**Table 4.9:** Gender classification accuracy for the different measurement sets (M0-M3) and age groups. MFCC feature classification results are shown for comparison.

| Age group | Baseline set M0 | Voice source measure sets M1 | M2 | M3 | MFCC features |
|---|---|---|---|---|---|
| 8-9 | 59.75% | 58.76% | 58.18% | 59.83% | 59.01% |
| 10-11 | 64.23% | 64.07% | 67.30% | 65.39% | 58.34% |
| 12-13 | 59.91% | 63.51% | 65.50% | 68.63% | 68.91% |
| 14-15 | 84.88% | 86.50% | 86.18% | 82.93% | 81.63% |
| 16-39 | 95.03% | 95.26% | 95.15% | 94.85% | 95.79% |

**Table 4.10:** Gender classification accuracy for age group 12-13, distinguishing between males and females.

| Set | M | F | Total |
|---|---|---|---|
| M0 | 59.28% | 60.60% | 59.91% |
| M1 | 63.24% | 63.80% | 63.51% |
| M2 | 63.06% | 68.20% | 65.50% |
| M3 | 66.67% | 70.00% | 68.63% |

using measure set M2. The SVM classifier performs comparably with the human subjects for the talkers aged 14 and above. For talkers aged below 14, the results are somewhat mixed and the accuracies reduce with decreasing age; however this trend also exists with the human classifiers. In effect, in the "total" section of the table, the AUT results agree well with the perception results.

Since the SVM was only given the target vowels, and the listeners were able to listen to the whole target word, it seemed only fair to see how listeners would perform when given only short vowel segments. Interestingly, for talkers of age 15 and above, the results were similar to gender classification using target word (about 90% recognition accuracy) and our experimental subjects were mostly using $F_0$ to do the classification. For talkers of age 14 and below however, our experimental subjects all agreed that their decisions on target vowels were mostly based on chance; the removal of the contextual information reduced the ability to distinguish between genders. As stated in [POA01]: "...prosodic features that are overlayed (suprasegmentals) upon sound segments in words, phrases, or sentences and include intonation, stress, duration, and juncture maybe important in gender identification."

### 4.4.4 Summary

In this work, the role of voice source measures in automatic gender recognition was examined and compared with the results of perceptual experiments performed on the same database. Vocal tract and voice source measures were extracted from a large database of 3880 utterances spoken by 205 males and 160 females. Formant frequencies and formant bandwidths were used as vocal tract measures, and $F_0$, $H_1^*-H_2^*$ (related to open quotient and asymmetry), and $H_1^*-A_3^*$ (i.e. spectral tilt) were used as voice source measures. The slopes (derivatives) were also calculated

**Table 4.11:** SVM gender classification accuracy, in percent, using measure set M2 compared with perception results from this paper (PER1) and from Perry et al. [POA01](PER2). Dashes indicate unavailable values. The perception experiments used the target words.

| Age | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|
| | Males | | | | | | | | |
| AUT | - | | - | | 67 | | 83 | | 94 |
| PER1 | 39 | | 72 | | 91 | | 100 | | 100 |
| PER2 | 74 | - | - | - | 82 | - | - | - | 99.7 |
| | Females | | | | | | | | |
| AUT | - | | - | | 68 | | 90 | | 97 |
| PER1 | 68 | | 75 | | 31 | | 70 | | 97 |
| PER2 | 56 | - | - | - | 56 | - | - | - | 95 |
| | Total | | | | | | | | |
| AUT | 58 | | 67 | | 66 | | 87 | | 95 |
| PER1 | 54 | | 73 | | 61 | | 86 | | 98 |
| PER2 | 65 | - | - | - | 69 | - | - | - | 97 |

for the voice source measures. Automatic gender classification using SVMs was performed on five age groups with different sets of acoustic measures.

Using a baseline measure set consisting of $F_0$, the first three formants ($F_1$, $F_2$, $F_3$) and the first two bandwidths ($B_1$, $B_2$), it was found that adding the two voice source measures $H_1^* - H_2^*$ and $H_1^* - A_3^*$ yielded the most consistent classification accuracy improvement over the baseline. For age group 8–9, the results were all below 60%, slightly higher than chance, however for ages greater than 9, using these two measures increased the classification accuracy, although the im-

provements decreased for older talkers as the role of $F_0$ became more dominant. The measure sets which included the slopes $\triangle F_0$ and $\triangle H_1^* - H_2^*$ did not produce consistent results and in some age groups actually reduced the classification accuracy.

Perception experiments using the target words showed similar results compared to the results of the SVM classifier, which used only the target vowel. Perception experiments using only the target vowel showed that for children aged 14 and below, classification accuracy was close to chance, suggesting that outside the vowel segment there exist suprasegmental cues, which could aid in automatic gender classification.

## 4.5 Application III: Prosody analysis

"Prosody" refers to properties of speech such as rhythm, timing, intonation, and stress. In American English, an important part of prosody relates to the prominence of a word within a phrase. This is usually marked by a pitch accent. Pitch accents, as a prosodic feature, allows a speaker to place contrastive stress on words within a phrase to indicate prominence or significance. Similarly, lexical stress allows a syllable to be more prominent than others within a word. Boundary tones signify groupings and allow a speaker to group words into intonational phrases and the choice of boundary tone can distinguish statements (Low or L–L%) from questions (High or H–H%). Accurate detection of pitch accents, stress, and boundary tones would benefit applications such as automatic speech recognition, speaker identification, and emotion classification.

With a few exceptions, previous studies of prosodic features have typically focused on the fundamental frequency ($F_0$), intensity, and duration. In [CHC05],

a large number of voice source related measures was analyzed using the Boston University Radio Corpus and it was found that there were no spectral harmonic measurements which could distinguish between accented and non-accented syllables. Similarly, [Oko06] found that correlates of pitch accents were: differences in peak fundamental frequency ($F_0$), peak intensity, and amplitude of voicing. In [SV96b], which studied Dutch speakers, and [Fan97], which studied Swedish sentences, it was found that stressed syllables are generally tenser, have more high frequency energy and lower open quotient of the glottal source. Since pitch-accented syllables are also stressed, it would be expected that these attributes might also apply to pitch-accented syllables. In [ISE06], it was found that these results were statistically significant if a distinction was made between low and high pitch accents. However, in that study, stressed syllables were compared with all other unstressed syllables in the corpus. When the effects of boundary-related tones were taken into account in later analysis, it was found that the results were only significant if the speakers were separated by gender.

In this study, using a prosodically-labeled corpus, which is carefully constructed to have the same words in different prosodic contexts, we examine how acoustic measures of lexical stress are affected by the presence of pitch accent, gender of the talker, and boundary tones. Acoustic measures were estimated and contours were fitted to these measures based on a weighted least squares error criterion. Analysis of variance (ANOVA) was performed to assess the statistical significance of the results.

### 4.5.1 Speech corpus

The corpus consists of data from [Eps02] along with new recordings of the same sentences so that the total number of speakers is 10: 5 males and 5 females. For

each speaker, 10 repetitions were recorded for each of the following sentences, where the bold word is accented:

- **Dagada** gave Bobby doodads.
- Dagada gave Bobby **doodads**.
- **Dagada** gave Bobby doodads?
- Dagada gave Bobby **doodads**?

The declarative and interrogative sentences induce the subjects to place contrasting boundary tones on the same word for the different sentence types.

Subjects were native speakers of Western American English between 21–35 years old. Signals were recorded in a sound-attenuated booth with a 1.0" Brüel & Kjær condenser microphone placed 5 cm from the subjects' lips. The signals were sampled at 20 kHz and downsampled to 10 kHz. The first and last repetitions of each sentence were discarded for the final analysis.

Two graduate students manually segmented the sentences and used the ToBI [SBP92] transcription standard to label the corpus. For this study, the high and low pitch accents, denoted by $H^*$ and $L^*$, respectively, and the high and low boundary-related tones, denoted by H–H% and L–L%, respectively, on the words "Dagada" and "doodads" were analyzed. Syllables with primary lexical stress as on "ga" in "Dagada" and on "doo" in "doodads" are underlined. For the analysis of "Dagada", 32 files from a male speaker who pronounced the word as "Dagada" were discarded, while for the "doodads" 10 files were discarded as the $F_0$ tracker did not provide reliable data. The final distribution of prosodic labels was 69/97/122 ($L^*/H^*/noPA$) occurrences for "Dagada" and 81/82/75/72 ($L^*$H–H%/$H^*$L–L%/L–L%/H–H%) occurrences for "doodads". Note that $noPA$

indicates no pitch accent and that the labels for "doo" can be L*/H* or none, while for "dads" they are either L–L% or H–H%.

### 4.5.2 Voice quality related measures

Three measures related to the voice source and voice quality were estimated: $F_0$, $H_1^* - H_2^*$, and $H_1^* - A_3^*$. These measures were estimated over the entire duration of each sentence using the `VoiceSauce` application (Section 4.2) with the following settings: $F_0$ was estimated using the STRAIGHT algorithm and formant frequencies were estimated with the Snack sound toolkit (window length of 25 ms, frame shift of 1 ms and pre-emphasis factor of 0.96).

### 4.5.3 Contour Fitting and Analysis

For each word, contours were fitted to the three voice source measures according to a weighted least squares error criterion based on the signal energy, $E(n)$. When the energy falls below a certain threshold, as would occur in-between syllables of a word, the voice source measures become less reliable, and hence, less weighting is applied to the error function. The error weighting function, $W(n)$, was determined by $E(n)$, with the threshold, $E_{th}$, at a quarter of the mean energy of the word. After $E(n)$ drops below the threshold, the weighting function decreases exponentially[3], as shown in Eq. 4.2.

$$W(n) = \begin{cases} 1, & E(n) \geq E_{th} \\ e^{-(E_{th} - E(n))/E_{th}}, & E(n) < E_{th} \end{cases} \tag{4.2}$$

The use of the error weighting function ensured that only the most reliable $F_0$ values were used for the contour fitting. Although raw $F_0$ values, in reality, are

[3]Other functions were tried, such as linear and piece-wise linear functions, but the exponential function provided the best performance due to its smooth roll-off.

not always continuous during a voiced stop such as the /d/ and /g/ in our target syllables (i.e., microprosody), the closure duration of a voiced stop is usually small compared to the vowel duration, allowing the contour mapping to effectively smooth over these regions.

Similar to what was done in [KGC05], weighted Legendre polynomials were used for the contour approximations due to their orthogonality property. Each Legendre polynomial, $P_i(n)$ is associated with a coefficient, $a_i$, which enables a data vector, $y(n)$, to be approximated as $y(n) \approx \sum_{i=0}^{N} a_i P_i(n)$, where $N$ is the desired polynomial order. The coefficients $a_i$ provide a simple way to approximate a data vector. For this study, we set $N = 3$ since the longest word in the test corpus consists of three syllables. Eq. 4.3 shows the error criterion, $E_a$, used in the optimization of the $a_i$'s.

$$E_a = \sum_n \Big( y(n) - \sum_{i=0}^{3} a_i P_i(n) \Big)^2 \cdot W(n) \tag{4.3}$$

The orthogonal property of Legendre polynomials enables each coefficient to be optimized separately. For simplicity, we used iterations of the intermediate value theorem to find the optimal $a_i$'s. Iterations were stopped when the $a_i$ values did not change within five decimal places. The four coefficients ($a_0$, $a_1$, $a_2$ and $a_3$) used in this study represent, respectively, the Legendre polynomials $P_0(x) = 1$ (related to the mean), $P_1(x) = x$ (related to linear slope), $P_2(x) = \frac{1}{2}(3x^2 - 1)$ (related to quadratic convexity/concavity), and $P_3(x) = \frac{1}{2}(5x^3 - 3x)$ (related to cubic behavior).

For each word, contours were fitted to the three voice source measures ($F_0$, $H_1^* - H_2^*$, and $H_1^* - A_3^*$) and the results were manually checked for all utterances; 29 $F_0$ contours at the beginning and the end of the utterances had to be manually corrected. For each prosodic event (H*, L*, H*L–L%, L*H–H%, H–H% and L–L%), the means of the coefficients were calculated, enabling a direct comparison

between the effects of each prosodic event. Two-way ANOVA tests, from the software package SPSS (v16.0) were then performed on the coefficients, with the fixed factors being speaker and prosodic feature. The $p$ (probability of null hypothesis) values, $F$ (ratio of the model mean square to the error mean square) values, and partial $\eta^2$ (measure of effect size) values are reported for some cases.

### 4.5.4 Results

#### 4.5.4.1 Pitch accent

For the word "Dagada", as expected, most talkers showed higher/lower $F_0$ values for H$^*$/L$^*$ pitch accented syllables compared to the unaccented (*noPA*) case. Figure 4.6 shows $F_0$ contours averaged over data from the male talkers for the unaccented and accented pronunciations of the word. Interestingly, for H$^*$, most talkers showed a minimum value close to the end of the first syllable ("Da") and a maximum value at the beginning of the last syllable ("da"), where the $F_0$ maximum was about 15 Hz higher for H$^*$ compared to *noPA*. That is, the $F_0$ maximum did not occur during the stressed "ga" syllable but was delayed to the beginning of the next syllable. The $F_0$ drop before the actual maximum indicates that these cases should perhaps be labeled with L+H$^*$, instead of H$^*$, although this distinction was sometimes difficult to make perceptually. Here, we consider both L+H$^*$ and H$^*$ to be of the same category. For the L$^*$ case, both genders showed an $F_0$ minimum at the middle of the stressed "ga" syllable, where it was about 15 Hz lower for L$^*$ compared to *noPA*. For 7 out of 9 talkers the delay between $F_0$ maximum for H$^*$ and $F_0$ minimum for L$^*$ was about 100 ms. For one female talker, there was no delay, and for another female talker, the delay was 200 ms. The delay may be due to the dip in $F_0$ before the H$^*$, which provides more contrast for the following high pitch accent. ANOVA results on the effects

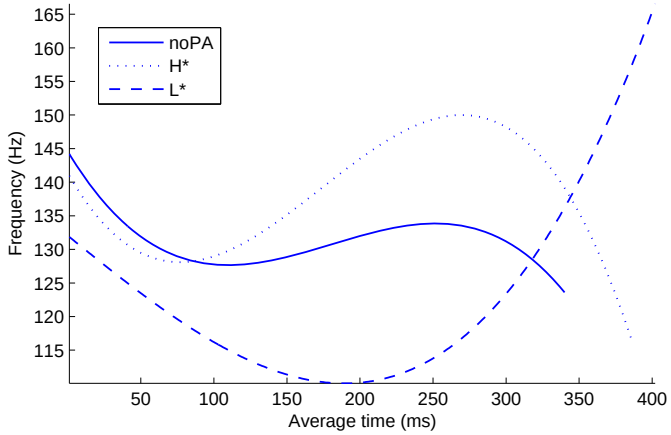



**Figure 4.6:** Average stylized $F_0$ contours "Dagada" (males).

of *noPA*, H*, and L* were significant for all speakers and all four polynomial coefficients.

Both genders also exhibited similar $F_0$ contours for the boundary word "doodads". Figure 4.7 shows $F_0$ contours for female talkers for each of the four prosodic events (L–L%, H–H%, H*L–L%, and L*H–H%). With few exceptions, the $F_0$ contour for H–H% increased monotonically ($a_1 > 0$), whereas for L–L% it decreased ($a_1 < 0$). For all talkers the contour for L–L% always lay below the contour for H–H% and the contours for accented words (L*H–H% and H*L–L%) lay mostly between the contours for L–L% and H–H%. The delayed $F_0$ peak for the H* case which was observed for "Dagada" was not as pronounced for "doodads", with only a slight delay observed for some talkers. This could be due to the influence of the boundary tone and/or due to the stress structure of the word. Interestingly, most speakers showed a slightly lower/higher $F_0$ before a high/low tone, respectively. This has also been observed in Mandarin [Xu97]. ANOVA analysis on the prosodic events showed that all four coefficients were statistically significant for both male and female speakers. As expected, both words show female $F_0$ contours with larger values and range than males (M/F:

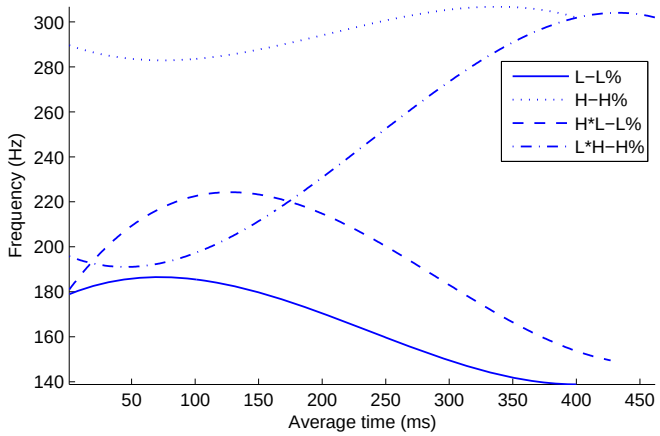



**Figure 4.7:** Average stylized $F_0$ contours for "doodads" (females).

110-155 Hz/190-260 Hz).

The duration of the word "Dagada" in accented cases was always longer compared to the unaccented cases and was confirmed with ANOVA analysis ($p/F/\eta^2$ = 0.00/139.7/0.52), which tested the significance of the durational change in "ga" with accentedness as a factor. The same trend was also found for "doodads", but with a smaller effect size (ANOVA: $p/F/\eta^2$ = 0.00/30.8/0.09). A similar result was reported in [TW99].

Interestingly, the acoustic measures $H_1^* - H_2^*$ and $H_1^* - A_3^*$ were not found to be distinctive for pitch accent. On average, $H_1^* - H_2^*$ exhibited the same convex shape regardless of the pitch accent type and $H_1^* - A_3^*$ showed some gender dependency. This may have been due to the influences of the lexical stress on the target word.

#### 4.5.4.2 Lexical Stress

In "Dagada", $H_1^* - H_2^*$ values seem to correlate well with lexical stress regardless of pitch accent. All talkers showed similar convex ($a_2 > 0$) contour shapes with a minimum during the stressed syllable "ga". Figure 4.8 shows these contours for

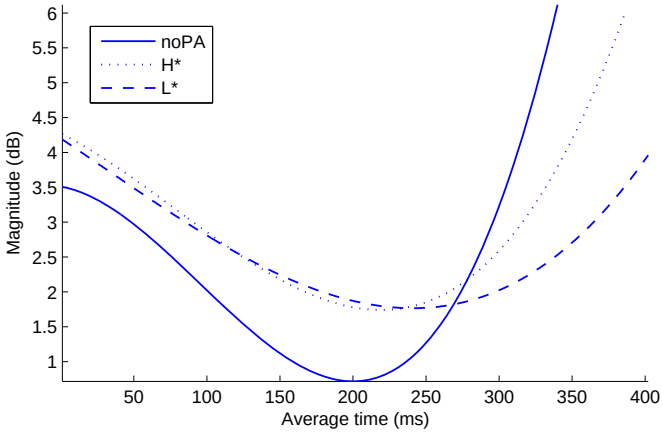



**Figure 4.8:** Average stylized $H_1^* - H_2^*$ contours for "Dagada" (males).

male talkers for each of the three prosodic events (*noPA*, H*, L*). For all talkers and independent of accentedness, $H_1^* - H_2^*$ was larger at the onset and the offset of the word than on the middle, stressed syllable "ga", possibly indicating a smaller open quotient and tenser voice quality for the stressed syllables. An ANOVA test on the raw $H_1^* - H_2^*$ mean values against the fixed factors speaker and syllable position within the word was significant with $p/F/\eta^2 = 0.00/68.17/0.15$. On average, the stressed syllable "ga" was about 2.5 dB and 4 dB lower than the surrounding syllables for males and females, respectively. As expected, "Dagada" $H_1^* - H_2^*$ contours showed higher mean values (M/F: 2.5 dB/4.9 dB) and a larger range (M/F: 0.5-5.5 dB/2-9 dB) for females when compared to male speakers [HC99].

As for "doodads", the results for $H_1^* - H_2^*$ contours seem to be speaker and gender dependent. On average, $H_1^* - H_2^*$ contours for L–L% lay above those of H–H% in female speech but the opposite was true for male speech. Contour minima and maxima could be found anywhere within the word and it was difficult to associate their locations with stress. This lack of consistency could be due to boundary tone effects.

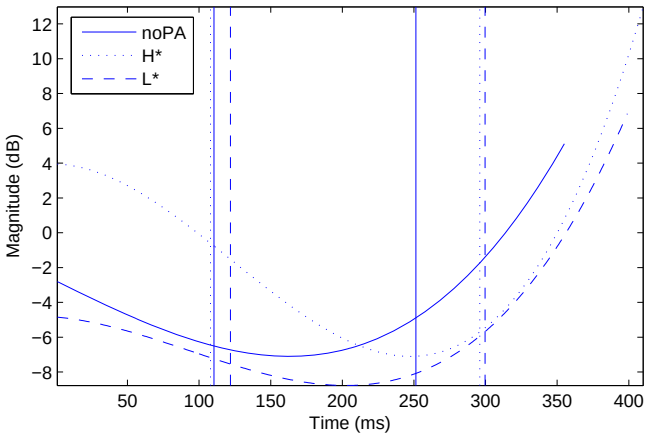



**Figure 4.9:** Stylized $H_1^* - A_3^*$ contours for "Dagada" for a male talker showing syllable boundaries for an instance of each prosodic case.

The $H_1^* - A_3^*$ contours appear to be gender dependent. For "Dagada", the average contours for both genders exhibited a parabolic shape. With the exception of one talker, male speech showed convex curves ($a_2 > 0$) for all three prosodic cases. For 3 out of the 5 female talkers, the opposite ($a_2 < 0$) was true for the accented cases. For almost all talkers the minima/maxima values occurred during the stressed "ga" syllable with male speakers showing a minimum for lexical stress regardless of pitch accent. Figure 4.9 shows these contours for one male subject. The figure also shows segment boundaries for the accented and unaccented cases. This indicates a more abrupt closure of the vocal folds on stressed syllables and agrees with [Oko06], [SV96b], [Fan97], and [ISE06]. As indicated earlier, for some female talkers, unaccented cases also had minima in "ga" but maximum values were observed when the stressed syllable was accented.

More consistency was found for the $H_1^* - A_3^*$ contours for "doodads" which, on average, had concave parabolic shapes. With the exception of two female talkers, the contours showed a low value of $H_1^* - A_3^*$ which increased to a maximum around

mid-word and then decreased at the end of the word (end of the utterance). This result again suggests that stressed syllables have lower spectral tilt (more high frequency energy) and agrees with previous work. Compared to declarative sentences (L–L% and H*L–L%), interrogative sentences (H–H% and L*H–H%) had, on average, a lower $H_1^* - A_3^*$ contour on the phrase-final syllable "doo"; a similar observation was made in [ISE06].

### 4.5.5 Summary

Not surprisingly, pitch accents were clearly marked by differences in $F_0$ contours. For "Dagada", averaged contours revealed that for both genders, the L* event caused the $F_0$ minima to occur at the middle of the accented syllable, while for the H* case, $F_0$ maxima appear towards the end of the accented syllable. This delayed peak, which was observed for almost all speakers for "Dagada" but not for "doodads", has implications for analyses which use mid-syllable values. For all speakers, the syllable and hence, word duration was longer for the accented cases than for non-accented cases.

For "Dagada", lexical stress was clearly marked by the convex shape of the $H_1^* - H_2^*$ contours which indicate a tenser voice (lower open quotient) on the stressed syllable; this measure seemed to be independent of pitch accent. However, this trend was not found for "doodads" possibly due to the influence of boundary tones. The spectral tilt measure ($H_1^* - A_3^*$) was seen to be gender dependent for "Dagada", with the contour decreasing for the stressed syllable for male speech, while for female speech, this was true only for the unaccented case. For "doodads", the boundary-related tone, especially H–H%, generally caused the $H_1^* - A_3^*$ contours to decrease towards the end of the word, denoting lower spectral tilt or an increase in high-frequency energy.

These results suggest that acoustic cues of lexical stress can be affected by the presence of a pitch accent, boundary tone, and in some cases, gender of the talker.

## 4.6 Summary and discussion

In this chapter, a new application, `VoiceSauce`, was introduced which simplified the process of calculating voice source related measures. This application was then used in three different scenarios: voice quality analysis, automatic gender classification and prosody analysis.

In voice quality analysis, the voice source parameters, as represented by the model-fitted glottal area waveforms, along with the voice source related measures were analyzed for correlations with the type of voice quality, the presence/absence of glottal gaps and the $F_0$ type. It was found that, on average, the parameter $OQ$ and the spectral tilt measures, $H_1 - A_2$ and $H_1 - A_3$ were affected by both voice quality and incomplete glottal closures. The asymmetry parameter $\alpha$, and the measures *CPP*, *HNR05*, *HNR15*, and *HNR25* were shown to be mainly affected by the voice quality, especially of the breathy type. This suggests the possibility of a link between $\alpha$ and the way noise is generated in breathy voices.

Automatic gender classification of speakers of varying age groups was found to improve with the addition of the voice source related measures $H_1^* - H_2^*$ and $H_1^* - A_3^*$. Since the $F_0$ and formant features dominate for adult speakers, the improvements were mainly seen for the age groups 10–11 and 12–13 year olds. Results were comparable with human listeners using the same data.

In the analysis of prosody, high pitch accented syllables were found to be marked differently by the $F_0$ contour than low pitch accented syllables. In high

pitch accents, the $F_0$ peaks were often found near the end of the accented syllable whereas low pitch accents usually had $F_0$ troughs around the middle of the accented syllable; this issue is analyzed in greater detail in Chapter 5. Lexical stress was found to be indicated by decreases in $H_1^* - H_2^*$ which suggests a tenser voice on the stressed syllable. $H_1^* - A_3^*$ was seen to be gender dependent, but for boundary related tones, the measure typically decreased towards the end of the word, indicating an increase in high-frequency energy.

These three applications showed the various uses of voice source related measures, which can be obtained with more ease than actual voice source parameters. However, although the work in this chapter has shed a little more light on the subject, relating these measures back to the voice source parameters still requires more research. The gender classification experiments show there are definite difference between young male and female speakers, and with more direct data, these differences could eventually be quantified. The prosody analysis revealed an interesting contextual effect and perhaps serves as a caution that the parts of interest in a target syllable may not necessarily be in the middle of that syllable.

# CHAPTER 5

# Acoustic Correlates of High and Low Nuclear Pitch Accents in American English

In Section 4.5, it was found that the $F_0$ maxima for high pitch-accented syllables were occurring, not at the middle of the accented syllable as expected, but towards the end of the syllable. In this chapter, a carefully designed speech corpus was recorded and analyzed in detail to unravel the reasons behind the peak-shifting. Apart from $F_0$, which is used to calculate most of the acoustic measures listed in Table 4.1, the duration and energy of each word were also examined. Results showed that the three measures could be affected by the occurrence of multiple prosodic events on a target word. This effect has been loosely named as *tonal-crowding* in previous literature and its effects are further clarified in this chapter.

## 5.1 Prosody and pitch accents

Prosody refers to the properties of speech such as rhythm, timing, intonation and stress[1]. In American English, as in many other languages, an essential role of linguistic prosody is to signal phrase-level prominence and phrasing, using tonal targets and other cues, such as duration (see [ST96] for a review). In the

[1]Docherty (1990) defines it in the following way: "Prosody or the melody of speech is the process used to alter the meaning (linguistic prosody) or emotional force (affective prosody) of a sentence. The components of prosody are rhythm, pitch, tone and stress and they are articulated by modulation of the acoustic correlates of prosody; frequency, duration and amplitude".

Autosegmental-Metrical (AM) approach to intonation ([Pie80, BP86, Lad08]), prominence is usually marked by a high or low Pitch Accent on the target word; phrasing is marked by a high or low Boundary Tone on the lengthened final syllable of the phrase; and an additional tonal element, a high or low Phrase Accent, controls the $F_0$ between the last pitch accent of a phrase and the boundary tone on the final syllable. The AM approach proposes a sparse string of such High and Low tonal targets, represented independently from the segmental/word string but associated with it, to define each Intonational Phrase.

A challenging issue in spoken prosody is the difficulty of specifying the acoustic correlates of these tonal elements. One problem relates to the height of the target in the $F_0$ space: these entities are usually defined in relational terms which are difficult to quantify absolutely. For example, the $F_0$ level associated with a high pitch accent for one speaker, while higher than a low pitch accent for the same speaker in the same context, might correspond to the $F_0$ level of a low accent for another speaker. Similarly, because the overall pitch range often declines during an utterance, a paragraph or a conversational turn ([Cha93, HP86]), a high pitch accent late in a constituent may actually have a lower $F_0$ than a low pitch accent that occurred earlier. Thus, it is difficult to specify a threshold for $F_0$, above which the target is a high pitch accent and below which the target is a low pitch accent, for all speakers or even for a single speaker. Another problem relates to the alignment of the target with the text in time. For example, there is good evidence that the alignment of the $F_0$ turning point associated with one tonal target can be influenced by the position and type of adjacent tonal targets ([SP90, ALM06, AG07]). Finally, in addition to questions about $F_0$ levels and text alignment, there are gaping holes in our understanding of other candidate acoustic correlates of tonal targets, although valuable work has been done on the parameters of intensity (e.g. [KGC05]), duration (e.g. Turk and colleagues) and

voice quality (e.g. [PT91, DSO96]). In this chapter, we examine a few of the acoustic correlates of two kinds of American English pitch accents, high (denoted by $H^*$) and low (denoted by $L^*$), and how the presence of adjacent phrase accents and boundary tones on the same word can affect these correlates.

This study was inspired by some puzzling results in earlier work (Section 4.5), i.e. a striking difference in the alignment of the $F_0$ peak associated with a $H^*$, in two different words which appeared in two different positions in the intonational phrase: *Dagada* gave Bobby doodads. That is, when the $H^*$ fell on the earlier target word, *Dagada*, its peak aligned near the end of its stressed syllable (-ga-), but when the $H^*$ fell on the later word, *doodads*, it aligned earlier in its stressed syllable (doo-). Because the two pitch accent contexts differed in several ways, it was not possible to determine which of several potential factors was responsible for the difference in peak alignment. For example, the two words differed in the quality of the stressed vowel, /a/ vs /u/, a factor that has been proposed by [JM07] to influence alignment. In that work, statistical analysis showed a correlation between vowel height and peak alignment, with high-vowel and low-vowel peak positions differing by approximately 11%. However, the corpus consisted of uncontrolled sentences selected from a newspaper corpus, which made it difficult to exclude other possible influences such as phrase position and stress as possible contributors to those results. Section 4.5's target words *Dagada* and *doodads* also differed in a number of other ways, which may have contributed to the observed alignment difference; they had different numbers of syllables, different positions of the main-stressed syllable in the word, and different positions of the word in the utterance. Finally, the two words differed in whether or not, in addition to the pitch accent, they also carried the boundary-related tones (phrase accent and boundary tone) for the intonational phrase. *Dagada*, which carried no boundary tone because it occurred early in the phrase, showed a later

$F_0$ peak alignment than *doodad*, which was the last word in the phrase and so did bear the boundary tone on its final syllable and the phrase accent before it. Because a number of studies have suggested that closely adjacent tonal targets may influence each other's alignment, in the configurational context described as tonal crowding ([SP90, ALM06]), we designed this follow-up study to determine whether tonal crowding could account for the observed differences in alignment reported. The expanded corpus was carefully designed to control for factors such as vowel context, while systematically varying the number of syllables, stress pattern and structural position of the target word in the intonational phrase (early vs. late, phrase-final vs. non-final). In this way, we sought to test the hypothesis that tonal crowding from a boundary tone can result in early location of the extremum of the $F_0$ excursion associated with a pitch accent, and to determine whether other factors such as word length, stress pattern and early vs. late (but not final) position in the phrase have an effect. In particular, we hypothesize that the $F_0$ peak for a $H^*$ accented syllable will be consistently located in or just after the accented vowel in a wide variety of contexts, i.e. for words located early or late (but not finally) in the phrase, and for words with various numbers of syllables and locations of main stress, because in these conditions there is no need to make room for boundary tone targets later in the same word. However, if the accented word is phrase-final, so that boundary tones occur in the same word, the speaker will realize the peak earlier in the accented syllable, as predicted by tonal crowding, and perhaps at a lower $F_0$ value as well. In addition, earlier results led us to hypothesize that the trough for a $L^*$ accented syllable may not show such a systematic move toward earlier alignment under conditions of crowding by immediately-following boundary tones.

In addition to the effects of the alignment of the $F_0$ contour with the spoken words and syllables, other acoustic features of accents, such as energy levels and

duration, might also be subject to tonal crowding effects. However, changes in energy measures (for example, in the presence of boundary tones) are difficult to predict. Previous studies, such as [SV96b], [SV96a], [KGC05] and [RH06], have shown that energy measures (typically using spectral balance, intensity or banded frequency content) are correlated with the presence of stress or pitch accents, and that these measures tend to rise with the occurrence of these prosodic events. Based on these findings, we hypothesize that energy values will be higher ([RH06]) for pitch-accented syllables than for unaccented syllables; however, this difference will be smaller if a boundary tone immediately follows the accented syllable, due to falling subglottal pressure at the end of an utterance ([Sli07]).

We also hypothesize, based on earlier work, that pitch-accented syllables will be longer ([TW99]) than unaccented syllables in the same position in the phrase, and that word-final syllables will also be longer when they occur in phrase-final position than when they are phrase-medial ([Kla76a, BE94, TS07]). The stimulus set is designed to test these findings from earlier work in a larger number of speakers, 10 male and 10 female. By including twenty participants in the experiment we hope to shed light on the generalizability of observations about the acoustic correlates of pitch accents in American English, both across gender and across individual speakers.

An additional pattern that was observed in Section 4.5 was that H* accents behave somewhat differently from L*s, and we include analyses of the same type in this more extensive study. For example, we will test the hypothesis that $F_0$ movements are relatively less extreme for L* than for H*, that the energy increase associated with pitch accents is less for L* than for H* even when the number of pitch periods included in the measure is controlled, and that L* accents do not exhibit the same property of shifting alignment of the $F_0$ trough under conditions

of tonal crowding as H* accents do for the peak.

## 5.2 Corpus and analysis methods

### 5.2.1 Corpus

The test corpus used in this chapter was carefully constructed to minimize or control for various factors, such as vowel type, syllable number and word position, which could influence the results. This corpus was designed in collaboration with professors Stefanie Shattuck-Hufnagel, Nanette Veilleux and Sun-Ah Jun. The corpus consists of spoken elicited utterances with specified pitch accent and boundary tone locations and types. The utterances are prosodic variations of the 2 sentences, *Dagada gave Anne a dada* and *A dada gave Anne dagadas*, with the target words being *dagada* and *dada*. For each utterance, a single pitch accent (H* or L*) is produced on either the early target word or the late one, in either a declarative setting or an interrogative setting. The nonsense words *dagada* and *dada* were used to ensure that the lexically-stressed syllables carrying the pitch accents (-ga- and da-, respectively) had the same vowel in all cases, avoiding any vowel-specific effects. The declarative and interrogative forms of the sentences were used to elicit the phrase-final tone sequences L–L% for H* utterances and H–H% for L* utterances, on the assumption that opposite polarity for the accent vs. the pitch accent and boundary-related tones would increase the chances of obtaining clearly-detectable $F_0$ peaks and troughs for the accents. The eight configurations of the sentences are listed below, with the accented target word in bold font.

- ***Dagada*** *gave Anne a dada.*

  H* L–L%

- *A dada gave Anne **dagadas**.*

  H* L–L%

- *A **dada** gave Anne dagadas.*

  H* L–L%

- *Dagada gave Anne a **dada**.*

  H* L–L%

- ***Dagada** gave Anne a dada?*

  L* H–H%

- *A dada gave Anne **dagadas**?*

  L* H–H%

- *A **dada** gave Anne dagadas?*

  L* H–H%

- *Dagada gave Anne a **dada**?*

  L* H–H%

These same eight sentences were also recorded with the unaccented word *daily* added at the end of the sentences, to carry the boundary tone; this allowed us to determine the effects on the pitch accent realization in the late target word when the boundary tone was moved to a following word. Each of these 16 sentences was elicited with a prompt question or statement, to ensure the correct placement of the tones. For example, to elicit a H* tone on the early target word *dagada*, and a L–L% boundary tone on the unaccented late target word *dada*:

Prompt: *Was it Dagada or Daguda that gave Anne a dada?*

Response: ***Dagada*** *gave Anne a dada.*

Recordings were made for 20 native speakers of American English (10 males/10 females) between the ages of 17 and 30. Some of the speakers were recorded under the supervision of Professor Shattuck-Hufnagel at MIT, while the remainder were recorded at the UCLA Department of Linguistics under the supervision of Professor Jun. Some of these speakers were phonetically trained; for the speakers who were not phonetically trained, their responses were checked for the correct placement of the pitch accents and boundary tones. For each sentence, 5 repetitions were recorded, for a total of 1600 utterances. The recordings were made in a sound booth at an effective sampling rate of 16 kHz. Manual segmentation of the target words *dagada* and *dada* from their context and of their main stress vowel were performed.

### 5.2.2 Analysis methods

Measures related to the $F_0$ contour, energy, and duration were estimated for the analysis of the target words and main-stress syllables. $F_0$ measures were extracted with the `VoiceSauce` application (Section 4.2) with a frame shift of 1 ms. Legendre polynomials of orders 3 to 7 were fitted to the $F_0$ values for each of the target word as in Section 4.5.3. Examples of such fits are shown in Figures 5.1 and 5.2 for high and low pitch accented target words, respectively.

The $F_0$ minimum and maximum values were calculated from the smoothed contours and normalized to each speaker's mean $F_0$ value ($\overline{F_0}$). $\overline{F_0}$ is the average $F_0$ over all of that speaker's utterances. The minima and maxima were expressed as a percentage of the speaker's $\overline{F_0}$. For example, the normalized value for a particular $F_0$ maximum ($F_{0max}$) is calculated as $(F_{0max} - \overline{F_0})/\overline{F_0} \times 100\%$.

Energy measures were calculated through `VoiceSauce` with a variable window

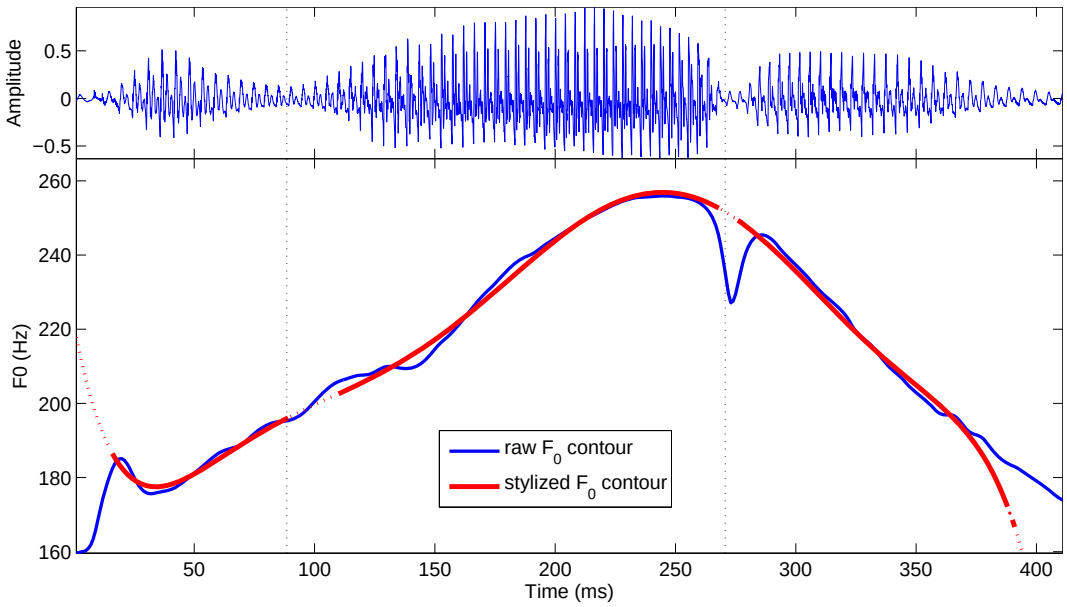



**Figure 5.1:** Example of polynomial fitting for the target word *dagada* with a high ($H^*$) pitch accent. The top panel shows the waveform, the bottom panel shows the raw and stylized $F_0$ contours. The dotted vertical lines mark the position of the manual segmentation.

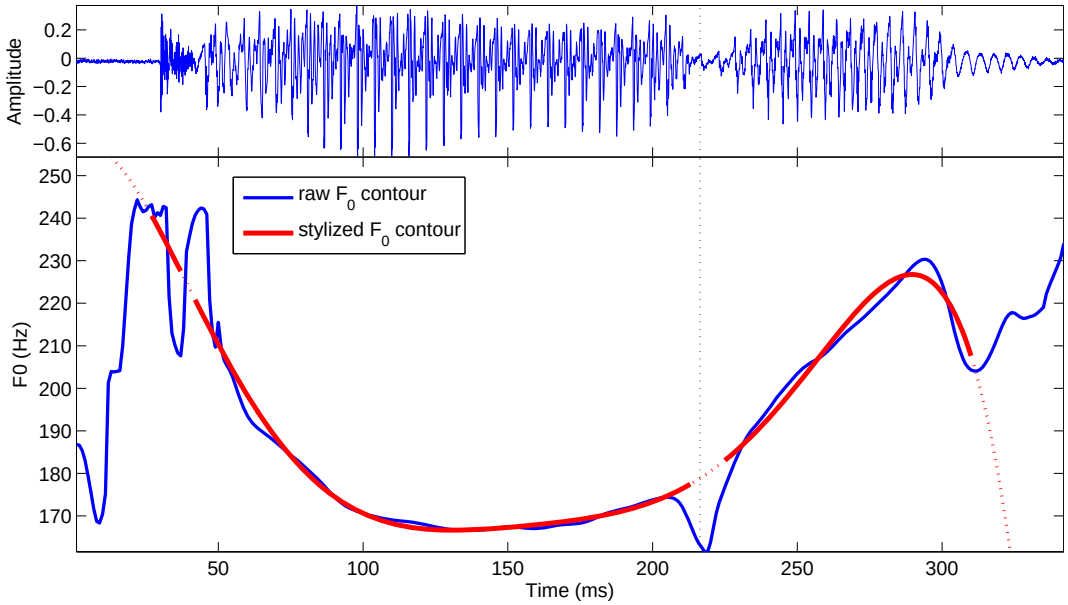



**Figure 5.2:** Example of polynomial fitting for the target word *dada* with a low (L*) pitch accent. The top panel shows the waveform, the bottom panel shows the raw and stylized $F_0$ contours. The dotted vertical line marks the position of the manual segmentation.

size set to 3 pitch periods, as determined by the STRAIGHT-estimated $F_0$ value at that point. For example, the $F_0$ value at the H* peak in Figure 5.1 has a value of 255 Hz; with a sampling frequency of 16 kHz, this leads to a window size of $\lfloor 16000/255 \times 3 \rfloor = 188$ samples. Utterances were normalized to have the same maximum energy value. The energy of each syllable is normalized with respect to the utterance's mean energy value and then used in ANOVA tests.

The duration of each main-stressed vowel was obtained from the manual segmentation. Onset and offset times were taken at the points where there was evidence of syllable closure or release, such as the abrupt fall in signal amplitude or the sudden loss of signal periodicity. These points are shown in Figures 5.1 and 5.2 as the vertical dotted lines.

The results of the analyses were grouped according to the gender of the speakers. This was done due to the well-known physiological and acoustical differences between male and female speakers ([Tit89]). Furthermore, it was shown in [ISA07] that many measures related to the voice source were dependent on the value of $F_0$ and thus, may be attributed to gender differences.

## 5.3 Results

We focused on the vowels in the main-stress syllables of the target words, which had relatively clear boundaries. We distinguish between several properties of the analyzed syllables: early vs. late position of the target word in the utterance; for late position, boundary vs. non-boundary position; position of the lexically-stressed syllable in the target word (medial in *dagada*, initial in *dada*); accentedness (accented vs. non-accented) and if accented, whether the accent was H* or L*. No vowel effects were examined because the same vowel /a/ was

intentionally used in all target syllables.

Four types of positions were examined, illustrated here with the labels *no-bnd-early*, *bnd*, *no-bnd-early-daily* and *no-bnd-late-daily* for declarative sentences with the target word, *da**ga**da*, where the -***ga***- syllable is always stressed and can be accented or not:

1. no-bnd-early: *Da**ga**da gave Anne a dada.*
2. bnd: *A dada gave Anne da**ga**das.*
3. no-bnd-early-daily: *Da**ga**da gave Anne a dada daily.*
4. no-bnd-late-daily: *A dada gave Anne da**ga**das daily.*

Additionally, in this section, for notation purposes, we will refer to all non-boundary (*no-bnd-early*, *no-bnd-early-daily* and *no-bnd-late-daily*) cases as *no-bnd* and all unaccented cases as *no-acc*.

Analysis of variance (ANOVA) tests were performed using the software package SPSS (v. 16.0) to check for the statistical significance of the results. The two fixed factors, speaker and accent tone ($H^*/L^*$), or speaker and boundary (yes/no, and if yes, H–H%/L–L%) were used to examine the effects on the measures. Because of the large number of tests done, tests where the null hypothesis has a probability of $p < 0.001$ were considered to be statistically significant.

We present the results based on our hypotheses made earlier, that is 1) tonal crowding affects the position and height of the $F_0$ maxima/minima of a pitch-accented syllable, and 2) phrase final lengthening is increased when the phrase final word also includes an accent. We also examined energy measures to determine the acoustic effects of pitch accents and boundary tones on this parameter.

### 5.3.1 $F_0$

To test the hypothesis that tonal crowding has an effect on the acoustic measures of pitch-accented syllables, we first analyzed the $F_0$ contours of pitch-accented syllables in words which contained a boundary tone (*bnd*) vs. those not containing a boundary tone (*no-bnd*). To separate out the effects of the serial position of the word, we then analyzed the $F_0$ measures for the *bnd* vs. *no-bnd-early(-daily)* cases and the *bnd* vs. *no-bnd-late-daily* cases.

Tables 5.1 and 5.2 show the results of ANOVA tests when $F_0$ peak/trough positions and heights are tested against the fixed factors *no-bnd* and *bnd* for the target word *dagada* and *dada* respectively. For statistically significant (s.s.) results, the $F$ ratio and partial $\eta^2$ (measure of effect size) values are also given.

**Table 5.1:** Position of the $F_0$ peak/trough as a percentage of the speaker's vowel duration. The results shown are averaged for the male and female speakers for the target words *dagada* and *dada*; standard deviation values are shown in parentheses. The statistical significance (s.s.) column shows the ANOVA results for *no-bnd* vs. *bnd*. For significant results, the $F$ (ratio of the model mean square to the error mean square) and $\eta^2$ (measure of the effect size) values are given.

| $F_0$ peak/trough position mean (std.) in % | | Males | | | Females | | |
|---|---|---|---|---|---|---|---|
| | | no-bnd | bnd | s.s. $F(1,180)/\eta^2$ | no-bnd | bnd | s.s. $F(1,180)/\eta^2$ |
| *dagada* | H* | 85(17) | 65(13) | **77.1/.300** | 92(15) | 69(13) | **222.8/.537** |
| | L* | 58(14) | 52(14) | No | 59(13) | 58(19) | No |
| *dada* | H* | 83(13) | 70(17) | **40.0/.188** | 90(15) | 68(13) | **139.6/.429** |
| | L* | 51(14) | 48(19) | No | 55(11) | 54(14) | No |

In Table 5.1, the values in the rows represent the mean $F_0$ peak/trough po-

**Table 5.2:** Height of the $F_0$ excursion as a percentage of the speaker's mean $F_0$. Average results are shown for the *no-bnd* vs. *bnd* conditions for the male and female speakers for the target words *dagada* and *dada*; standard deviation values are shown in parentheses.

| $\triangle F_0$ mean (std.) in % | | Males | | | Females | | |
|---|---|---|---|---|---|---|---|
| | | no-bnd | bnd | s.s. $F(1,180)/\eta^2$ | no-bnd | bnd | s.s. $F(1,180)/\eta^2$ |
| *dagada* | H* | 24(20) | 8(17) | **78.7/.305** | 35(27) | 23(28) | **46.3/.194** |
| | L* | –26(12) | –29(12) | No | –35(9) | –40(16) | No |
| *dada* | H* | 21(19) | 8(17) | **41.2/.193** | 36(25) | 23(28) | **41.1/.181** |
| | L* | –27(12) | –29(11) | No | –35(8) | –42(11) | **28.6/.137** |

sitions relative to the duration of the target vowel, where 50% corresponds to the middle of the target vowel. For example, for male speakers, the H* accented *dagada* for the cases where the target word did not also include the boundary tone (i.e. the *no-bnd* case) had an $F_0$ peak which occurred, on average, at a position which was 85% of the mean vowel duration with a standard deviation of 17%. This is in contrast to the case where the target word also carried the boundary tone (*bnd*), which had a position of 65% of the vowel duration. No statistically significant effects were observed for L* accented target words. The results are similar for the target word *dada*, again showing the shift of the peak to an earlier position in the presence of the boundary tone.

In Table 5.2, the values in the rows denote the mean $F_0$ peak/trough excursion relative to the mean of the speaker's $F_0$ values in percent; i.e. 0% corresponds to a peak/trough exactly at the speaker's mean $F_0$ calculated over all the speaker's utterances. For example, for the female speakers producing *dada* in the non-

boundary condition (*no-bnd*), the H* accent resulted in an $F_0$ peak which was on average 36% higher than the speaker's mean $F_0$, while for the boundary case (*bnd*), the $F_0$ peak was only 23% higher, showing a lesser excursion of $F_0$ for the pitch accent in the presence of a boundary tone on the same word. Note that the boundary tone for a H*/L* accented target word is L–L% and H–H% respectively.

The ANOVA tests show that regardless of gender, for the H* accented vowels the *no-bnd*/*bnd* factors have a statistically significant effect on both the position and height of the $F_0$ peak. For both target words, the *no-bnd* case had an $F_0$ peak position which occurred much later than the *bnd* case and peak heights that were greater. Interestingly, the L* accented vowels showed no statistically significant shift in height or duration, with the exception of *dada* for female speakers for the *height* factor, where a weak effect size ($\eta^2 = 0.137$) is observed.

Since the *no-bnd* case contains instances where the pitch-accented word is at the start of the sentence (*no-bnd-early* and *no-bnd-early-daily*) and near the end of the sentence (*no-bnd-late-daily*), it is possible that the serial position of the word may also affect the results, although this was not hypothesized. For example, if $F_0$ declination occurs over the course of the utterance, the $F_0$ peak for a H* may be lower for an accent that occurs late in the utterance than for an accent that occurs early. To confirm that tonal crowding rather than declination is the main cause of the lower $F_0$ peak, we also analyzed the $F_0$ contour peak and height by comparing *bnd* vs. *no-bnd-early(-daily)* and *bnd* vs. *no-bnd-late-daily*. Figures 5.3 and 5.4 show the $F_0$ peak positions for H* for a typical male/female speaker producing dagada/dada respectively in the three different contexts; note that the *no-bnd-early* and *no-bnd-early-daily* cases showed similar results, hence were considered together. For these two speakers, it can be seen that the *bnd* cases generally had earlier $F_0$ peaks compared to both of the other two cases,

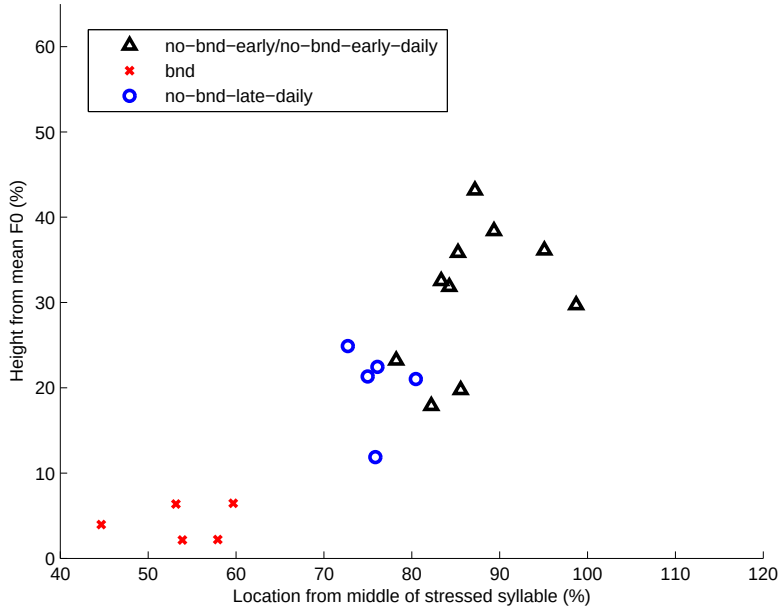



**Figure 5.3:** Scatter plot for the target word *dagada* showing relative $F_0$ peaks for H* and their relative positions in the accented target vowel for a male speaker in three different contexts: 1. no-bnd-early/no-bnd-early-daily (triangles); 2. bnd (crosses); 3. no-bnd-late-daily (circles).

indicating the temporal crowding effect of the boundary tone. The male speaker (Figure 5.3) also had $F_0$ peaks that were significantly lower for the *bnd* case than the other two cases, while for the female speaker (Figure 5.4) the $F_0$ peaks for the *bnd* case was lower than the *no-bnd-early(-daily)* case and similar to the *no-bnd-late-daily* case, indicating the possible compression effects of the boundary tone. These trends were observed for 8/10 male speakers and 9/10 female speakers. Note that the ANOVA tests in this section were carried out using the results from all speakers, including those that did not conform to the general trend.

Statistical analysis comparing $F_0$ peak/trough position and height for the *bnd* vs. *no-bnd-early(-daily)* conditions are shown in Tables 5.3 and 5.4 respec-

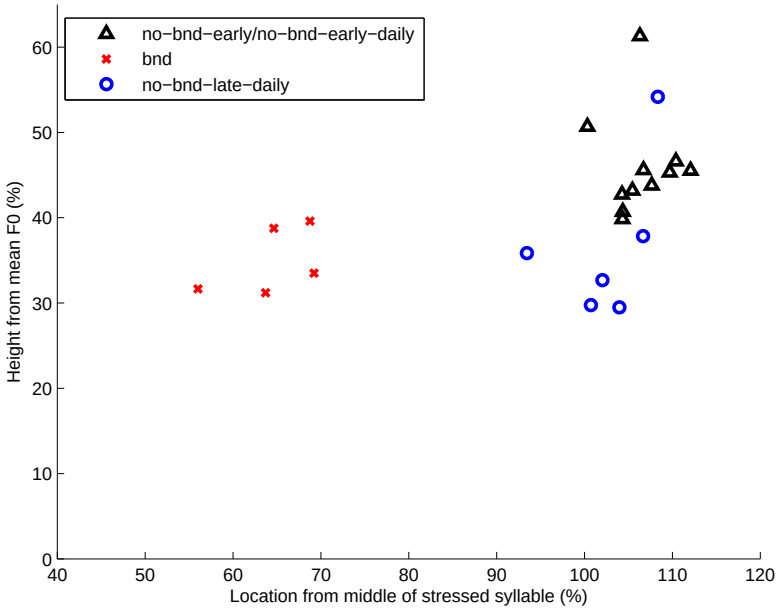



**Figure 5.4:** Scatter plot for the target word *dada* showing relative $F_0$ peaks for H* and their relative positions in the accented target vowel for a female speaker in the three different contexts.

tively, for target words *dagada*/*dada*. Similar to the trends shown in Tables 5.1 and 5.2 for the *bnd*/*no-bnd* comparison, it can be seen that for the H* pitch accent, the position and height of the $F_0$ peak are affected significantly by the presence/absence of the boundary tone, with the *no-bnd-early(-daily)* peak occurring later and higher than the *bnd* case. For example, for female speakers the H* accented target word *dagada* showed a peak position difference of 27% (from 69% to 96%) and a height difference of 18% (from 23% to 41%) when comparing the *bnd* case to the *no-bnd-early(-daily)* cases. L* accented words did not exhibit any statistical significance for the position measure, except for male speakers for the target word *dagada* which showed a small effect size ($\eta^2 = 0.104$). Similarly, although the height measure differences for L* accented target words are statistically significant, the difference between the means for the *bnd* and *no-bnd-early(-daily)* conditions are small with a relatively weak effect size.

**Table 5.3:** Position of the $F_0$ peak/trough as a percentage of the speaker's target vowel duration. Results shown are average values for the male and female speakers for target words *dagada* and *dada* in the *no-bnd-early(-daily)* vs. *bnd* condition; standard deviation values are shown in parentheses.

| $F_0$ peak/trough position mean (std.) in % | | Males | | | Females | | |
|---|---|---|---|---|---|---|---|
| | | no-bnd-early (-daily) | bnd | s.s. $F(1,270)/\eta^2$ | no-bnd-early (-daily) | bnd | s.s. $F(1,270)/\eta^2$ |
| *dagada* | H* | 85(17) | 65(13) | **167.3/.565** | 96(15) | 69(13) | **292.1/.678** |
| | L* | 59(14) | 52(14) | **14.9/.104** | 61(13) | 58(19) | No |
| *dada* | H* | 84(11) | 70(17) | **36.0/.229** | 93(15) | 68(13) | **171.1/.563** |
| | L* | 52(13) | 48(19) | No | 55(9) | 54(14) | No |

**Table 5.4:** Height of the $F_0$ excursion as a percentage of the speaker's mean $F_0$. Results shown are for the *no-bnd-early(-daily)* vs. *bnd* conditions; standard deviation values are shown in parentheses.

| $\triangle F_0$ mean (std.) in % | | Males | | | Females | | |
|---|---|---|---|---|---|---|---|
| | | no-bnd-early (-daily) | bnd | s.s. $F(1,270)/\eta^2$ | no-bnd-early (-daily) | bnd | s.s. $F(1,270)/\eta^2$ |
| *dagada* | H* | 26(19) | 8(17) | **71.6/.372** | 41(22) | 23(28) | **113.8/.461** |
| | L* | –24(11) | –29(12) | **32.8/.204** | –34(7) | –40(16) | **19.3/.130** |
| *dada* | H* | 26(19) | 8(17) | **71.6/.372** | 41(22) | 23(28) | **113.8/.461** |
| | L* | –26(12) | –29(11) | **45.8/.259** | –35(8) | –42(11) | **23.6/.153** |

Tables 5.5 and 5.6 show the $F_0$ results when the target words are tested for the *no-bnd-late-daily* vs. *bnd* effect for the position and height measures, respectively. Similar to the previous results, it can be seen that, regardless of gender and target word, the H* accented $F_0$ peak, on average, occurred much later for the *no-bnd-late-daily* case and these results were statistically significant. Height differences were less consistent than position differences, with only the male speakers showing statistical significance for the target word *dagada*. For L* accented syllables, only female speakers for the height measure on the target word dada showed any statistical significance, and the effect size is relatively weak.

These results are consistent with our hypothesis that the $F_0$ peak for an H* accent is usually located towards the end or just after the accented vowel, unless the need to realize other tonal targets (such as a boundary tone) on the same word causes the speaker to realize the peak earlier in the accented syllable, presumably in order to make room for the realization of the additional targets. This further

**Table 5.5:** Relative position of the $F_0$ peak/trough as a percentage of the speaker's target vowel duration. Results shown are average vales for the male and female speakers for target words *dagada* and *dada* in the *no-bnd-late-daily* vs. *bnd* condition; standard deviation values are shown in parentheses.

| $F_0$ peak/trough position mean (std.) in % | | Males | | | Females | | |
|---|---|---|---|---|---|---|---|
| | | no-bnd-late-daily | bnd | s.s. $F(1,183)/\eta^2$ | no-bnd-late-daily | bnd | s.s. $F(1,183)/\eta^2$ |
| *dagada* | $H^*$ | 80(17) | 65(13) | **55.5/.404** | 86(13) | 69(13) | **202.2/.697** |
| | $L^*$ | 55(13) | 53(14) | No | 56(13) | 58(19) | No |
| *dada* | $H^*$ | 80(15) | 70(17) | **11.6/.125** | 83(14) | 68(13) | **117.6/.586** |
| | $L^*$ | 48(14) | 48(19) | No | 54(13) | 54(14) | No |

supports the hypothesis that the location of the $F_0$ peak for an $H^*$ is affected more by tonal crowding ([ALM06]) from boundary tones than by the mere serial position of the accented word, or by its number of syllables. Interestingly, the $F_0$ peak position on average is later for the female speakers than for the males; this may be due to larger relative change in $F_0$ which could require more time to achieve. The lack of a consistent trend for $L^*$ pitch accented syllables is as predicted, and suggests the implementation of $L^*$ accents may be governed by different principles from those governing $H^*$ accents. These results also highlight the fact that cues which are found to be correlated with $H^*$ pitch accents may not necessary be correlated with $L^*$ accents; so that it is important to separate pitch accents into $H^*$ and $L^*$ categories when analyzing their correlates.

The results for the height measure for the $H^*$ accented syllables were seen to be more consistent for utterances which had the accented syllable at the start

**Table 5.6:** Relative height of the $F_0$ excursion as a percentage of the speaker's mean $F_0$. Results shown are for *no-bnd-late-daily* vs. *bnd*; standard deviation values are shown in parentheses.

| $\triangle F_0$ mean (std.) in % | | Males | | | Females | | |
|---|---|---|---|---|---|---|---|
| | | no-bnd-late-daily | bnd | s.s. $F(1,183)/\eta^2$ | no-bnd-late-daily | bnd | s.s. $F(1,183)/\eta^2$ |
| *dagada* | H* | 19(23) | 8(17) | **49.4/.376** | 24(29) | 23(28) | No |
| | L* | –29(11) | –29(12) | No | –38(11) | –40(16) | No |
| *dada* | H* | 13(16) | 8(17) | No | 26(27) | 23(28) | No |
| | L* | –30(11) | –29(11) | No | –36(8) | –42(11) | **15.2/.159** |

of the sentence (i.e. *no-bnd-early(-daily)*). For these cases, the $F_0$ peak was consistently higher than for the boundary cases. Although this trend was also seen for the utterances which had the late accented syllable (*no-bnd-late-daily*), those results were not statistically significant. This may be an effect of the need to realize the final low boundary tone (L–L%) on the word *daily*, which immediately follows the pitch-accented word. Another possible explanation is that the height of the $F_0$ peak for the late accented syllable may be influenced by the natural $F_0$ declination which can occur over the course of declarative statements.

### 5.3.2 Energy

We hypothesized that the energy change for a pitch accent might be similar across gender and pitch accent type, but different in boundary vs. non-boundary conditions, because of respiratory and subglottal pressure changes at the end of an utterance ([Sli07]). Figures 5.5 and 5.6 show, respectively, the scatter plots

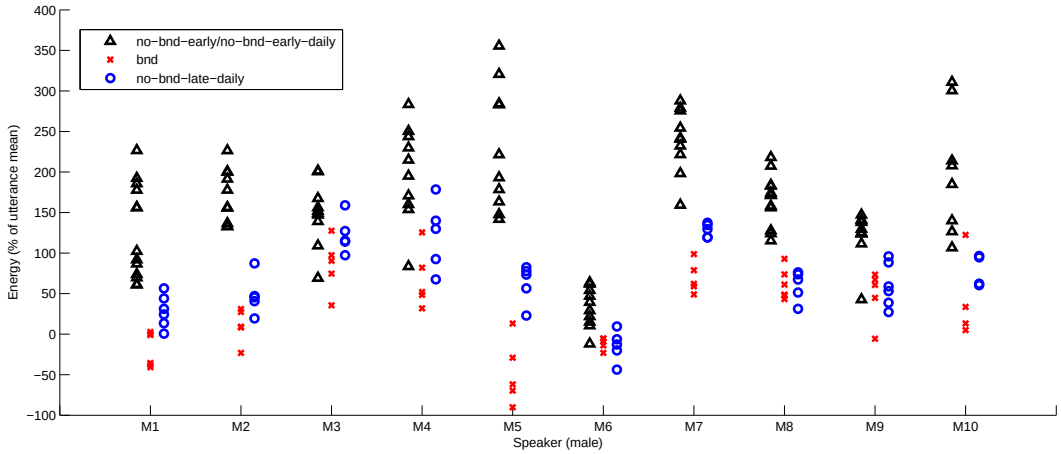



**Figure 5.5:** Scatter plot of relative mean vowel energy of the H* accented vowel for the target word *dagada* for all male speakers in three different contexts: 1. *no-bnd-early/no-bnd-early-daily* (triangles); 2. *bnd* (crosses); 3. *no-bnd-late-daily* (circles).

for male/female speakers of the relative mean energies for the H* accented vowel where the target word is *dagada/dada* in the three conditions, *no-bnd-early(-daily)*, *bnd*, and *no-bnd-late-daily*. The y-axis represents the change in relative energy in relation to the average energy of each utterance in percent, so a 100% value would correspond to twice the average utterance energy. It can be seen that for most speakers the boundary (*bnd*) case provides, on average, the least change in energy compared with the non-boundary cases (*no-bnd-early*, *no-bnd-early-daily* and *no-bnd-late-daily*). The considerable overlap between the *no-bnd-daily* and the *bnd* cases suggest that, for some speakers, the fall in energy could occur as early as the accented syllable of a penultimate word.

Two-way ANOVA results for the energy measure separated by gender, pitch accent tone-type and presence of adjacent boundary tones are shown in Tables 5.7 and 5.8 for target word *dagada* and *dada*, respectively. The values represent the percentage of change of the mean target syllable energy from the mean utterance

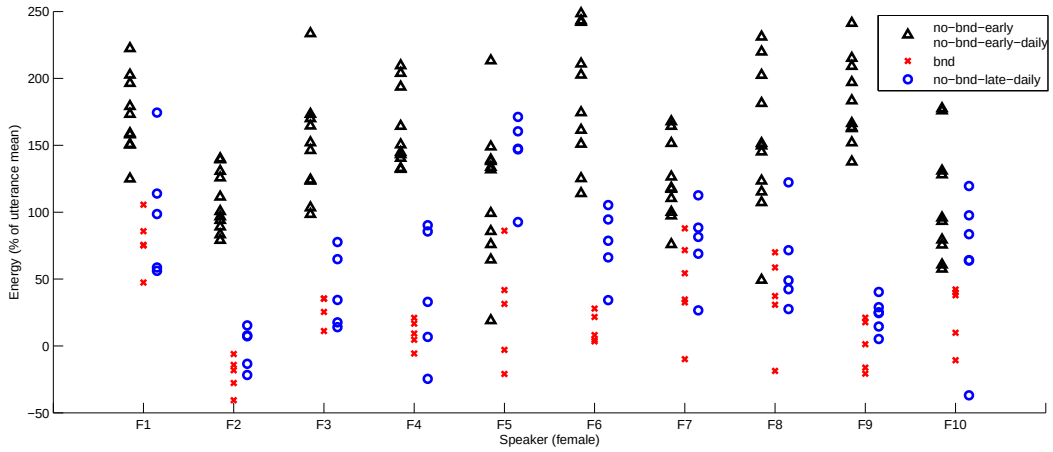



**Figure 5.6:** Scatterplot of relative mean vowel energy of the H* accented vowel for the target word dada for all female speakers in three different contexts: 1. *no-bnd-early*/*no-bnd-early-daily* (triangles); 2. bnd (crosses); 3. *no-bnd-late-daily* (circles).

energy. For example, the first row in Table 5.7 shows that the average energy of the H* accented vowel for male speakers in the *no-bnd*/*bnd* case is 131/30% higher than the mean energy of the utterance, statistically significant with an $F$-value of 114.2 and effect size of 0.382. It can be seen that on average, the change in energy was lower for the *bnd* case than for the *no-bnd* case regardless of gender and pitch accent. This trend also extends to the non-accented cases (*no-acc*) for both the interrogative (H–H%) and declarative (L–L%) utterances. Interestingly, L* accented syllables had a lower energy change than H* accented syllables and for female speakers, the energy change, on average, was significantly lower than even the *no-acc* cases. There were also significant differences associated with pitch accent type; for example, energy values generally showed greater variance for males across boundary conditions, and values were higher for H*/*no-acc*(H–H%) than for L*/*no-acc*(L–L%) regardless of gender. Thus, as with the $F_0$ peak location, the results for energy show the importance of context information in

understanding the acoustic correlates of tonal targets.

**Table 5.7:** Relative energy mean, standard deviation (std.) in parenthesis, of stressed syllables for the target word *dagada* for male and female speakers. Results are shown for *no-bnd* vs. *bnd*.

| Energy mean (std.) | Males | | | Females | | |
|---|---|---|---|---|---|---|
| | no-bnd | bnd | s.s. $F(1,185)/\eta^2$ | no-bnd | bnd | s.s. $F(1,185)/\eta^2$ |
| H* | 131(80) | 30(51) | **114.2/.382** | 125(74) | 9(39) | **132.9/.409** |
| L* | 36(72) | –10(49) | **39.1/.172** | –11(57) | –54(34) | **31.4/.147** |
| no-acc(H–H%) | 36(52) | –17(38) | **50.9/.215** | 35(48) | 13(40) | **14.2/.073** |
| no-acc(L–L%) | 3(61) | –67(26) | **68.6/.271** | 4(65) | –79(13) | **94.9/.338** |

### 5.3.3 Duration - effects of pitch accent on phrase final lengthening

Predictions about the effects of pitch accent on phrase-final lengthening are complex, since several different factors are at work. These include duration lengthening associated with main lexical stress ([BE94]), with the main-stress syllable of the phrase-final word ([TS07]), with the final syllable of the phrase ([Kla76b]) and with a pitch accent ([TW99]). We hypothesized that phrase-final lengthening might increase with the addition of a pitch accent on the phrase-final word, possibly in order to allow the speaker more time to achieve both prosodic targets. In this section, we first test the effects of high and low pitch accents on duration and target word position. We then test the phrase-final-lengthening effects in our corpus by comparing the final syllable durations for the non-accented words *dagada* and *dada* in the *late* vs. *early* vs. *bnd* conditions; the late case is where the unaccented word is positioned before the word *daily*, the *early* case

**Table 5.8:** Relative energy mean, standard deviation (std.) in parenthesis, of stressed syllables for the target word dada for male and female speakers. Results are shown for *no-bnd* vs. *bnd*.

| Energy mean (std.) | Males | | | Females | | |
|---|---|---|---|---|---|---|
| | no-bnd | bnd | s.s. $F(1,185)/\eta^2$ | no-bnd | bnd | s.s. $F(1,185)/\eta^2$ |
| H* | 140(68) | 38(42) | **178.0/.490** | 116(63) | 24(34) | **108.7/.369** |
| L* | 34(61) | –12(45) | **49.4/.210** | –30(49) | –50(32) | **16.0/.081** |
| no-acc(H–H%) | 32(58) | –3(50) | **20.6/.099** | 31(46) | 14(36) | No |
| no-acc(L–L%) | 5(60) | –66(32) | **71.1/.278** | 5(65) | –75(17) | **91.0/.322** |

is where the unaccented word appears at the start of the utterance and the *bnd* case is where the unaccented word is the phrase-final word. The main-stress syllable lengthening effect of the phrase-final word is then tested by comparing the stressed-syllable durations of the non-accented *dagada* and *dada* in the *late* vs. *early* vs. *bnd* condition. Finally, the hypothesized extra phrase-final lengthening effect is checked by comparing the final-syllable durations of the target words for the *no-bnd-early* vs. *bnd* conditions.

Figure 5.7 shows the average durations, for male and female speakers, of the main-stressed syllable (-ga-) for the target word *dagada* in the early, late and boundary conditions. It can be seen that regardless of word position, the average duration for the *Non* case is much lower than for either L* or H* cases, confirming the durational lengthening associated with pitch-accents ([TW99]). While ANOVA tests on the tone-type showed that the results are statistically significant, post-hoc analyses revealed that the L* and H* accented durations are virtually indistinguishable. This result means that while durational lengthening

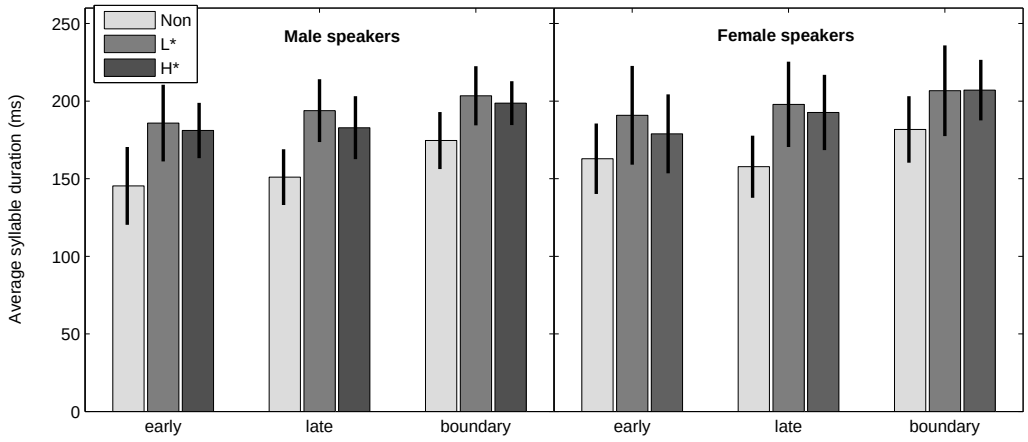



**Figure 5.7:** Average main-stressed syllable duration with no (Non), $L^*$, and $H^*$ pitch accents for male and female speakers for the target word *dagada* in the early, late and boundary positions.

occurs in the presence of pitch-accents, it is not affected by the pitch-accent type. Furthermore, the results also show that the boundary position has on average longer durations for the Non, $L^*$ and $H^*$ cases than for both the early and late positions. Results for the target word *dada* are similar.

Figure 5.8 shows the comparisons of the average durations and the corresponding error bars of the final syllable (*daga**da*** and *da**da***) for male and female speakers. The effects of phrase final lengthening can be clearly seen in the longer average duration for the *boundary* case. This result is statically significant for both words when the fixed factors, speaker and word position, are used in an ANOVA test; for male speakers, $F(2/2{,}381/383) = 413/414$, $p = 0.000/0.000$, and $\eta^2 = 0.68/0.68$ and for female speakers, $F(2/2{,}377/384) = 489/486$, $p = 0.000/0.000$ and $\eta^2 = 0.72/0.72$ for the unaccented words *dagada*/*dada*. Note that the two no-boundary cases, *early* and *late*, are not significantly different in their averaged final-syllable durations, indicating that for the *late* case, the

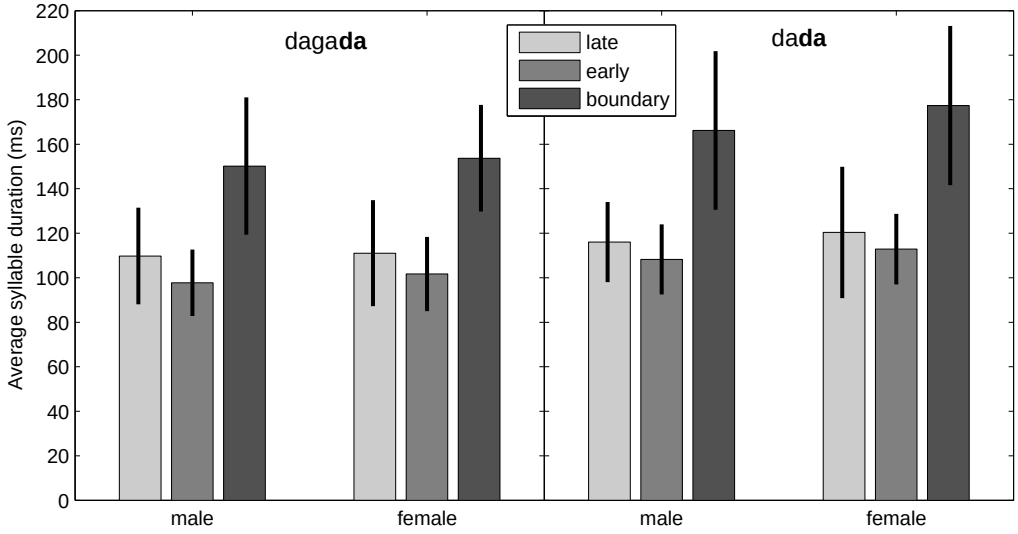



**Figure 5.8:** Average final vowel durations and error bars of the unaccented words *dagada* and *dada* for male and female speakers in the *late*, *early* and *boundary* positions. The increased duration for the *boundary* case confirms the effects of phrase final lengthening.

final vowel may not be close enough to the boundary for boundary-related effects to appear. This may be because it is too far away in time or because of the intervening word boundary.

Figure 5.9 shows the comparisons of the average durations and the corresponding error bars of the main-stressed syllable (*da**ga**da* and ***da**da*) for male and female speakers, when the target words are unaccented and in the *late*, *early* and *boundary* conditions. It can be seen that the boundary case has the largest average duration for both target words and for both genders. These results were all statistically significant ($p < 0.001$) and confirm the lengthening effect of boundaries on unaccented main-stress syllables reported in [TS07].

Figures 5.10 and 5.11 show a comparison of the average durations of the final vowel for male and female speakers for the target word *dagada* and *dada*

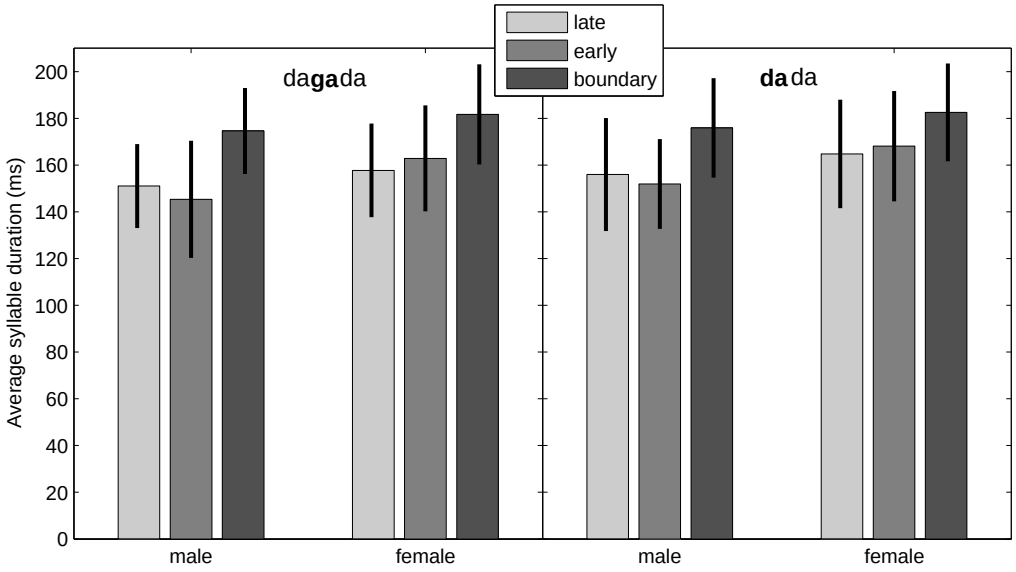



**Figure 5.9:** Average main-stressed vowel durations and error bars of the unaccented words *dagada* and *dada* for male and female speakers in the late, early and boundary positions. The increased duration for the boundary case confirms the unaccented main-stressed syllable lengthening at the boundary condition.

respectively, for three conditions: no accent on the preceding syllable, H* accent on the preceding syllable, and L* accent on the preceding syllable. On average, the duration of the final vowel increased when there was a preceding pitch accent. However, when the pitch accent was further categorized into H* and L* cases, it can be seen that there are some speakers who differ from the general trend; for example, speaker M1 and speaker F5 showed a shorter duration for the H*-preceding condition than for the no-accent condition. Statistical significance tests of the phrase-final vowel durations against the presence/absence of accent (and if present, H*/L*) confirm the general trends seen in the figures; the ANOVA results are shown in Table 5.9 and the results were statistically significant for both genders and target words. Interestingly, the mean durations show that on average, the L*-preceding case was marginally longer than the H*-preceding case and much longer than the no-accent case. This result indicates that while there is a slight difference between the L*-preceding and H*-preceding cases, the extra phrase-final lengthening is more dependent on the presence/absence of a preceding accent than on the type of accent.

## 5.4 Discussion

Our goals in this work were to test the hypothesis that tonal crowding contributed to the striking difference in the alignment of the $F_0$ peak associated with H* in phrase-final vs non-final words in Section 4.5, and to explore other aspects of the acoustic correlates of American English H* and L* pitch accents. In this section we first review the implications of our results for these goals (Section 5.4.1), then discuss the additional insights that emerge from access to results for 20 individual speakers (Section 5.4.2), and finally discuss in more general terms the concept of tonal crowding (Section 5.4.3): how this term has been used, the types of

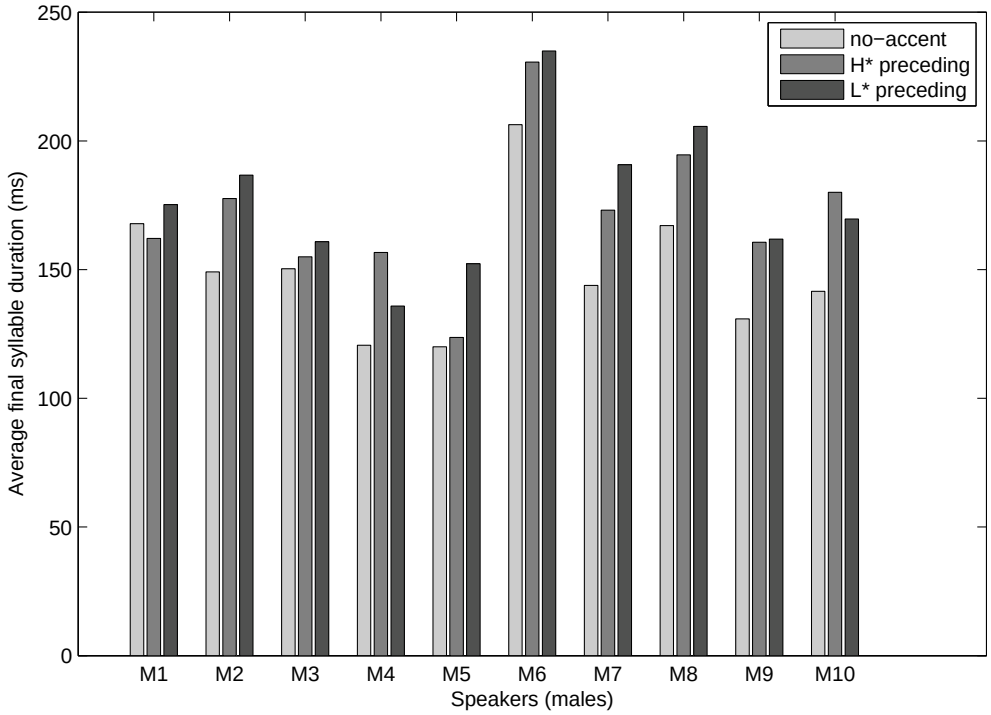



**Figure 5.10:** Average final syllable durations for male speakers for phrase final target word *dagada* with no preceding accents, with a preceding H* accent, and with a preceding L* accent.

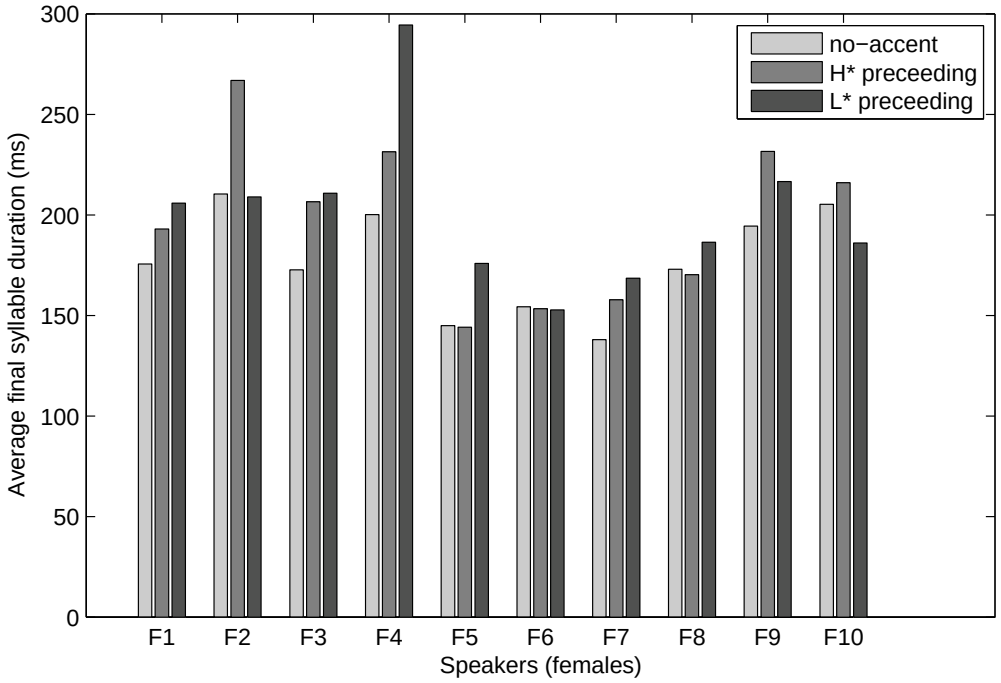



**Figure 5.11:** Average final syllable durations for female speakers for phrase final target word *dada* with no preceding accents, with a preceding $H^*$ accent, and with a preceding $L^*$ accent.

**Table 5.9:** Average duration, standard deviation (in parenthesis) of the final syllable of *dagada*/*dada* with no preceding pitch accent, with a H* preceding accent, and with a L* preceding accent. All results were statistically significant.

| Mean duration (ms) | Males | | | |
|---|---|---|---|---|
| | no accent | H* preceding | L* preceding | s.s. $F(2,174)/\eta^2$ |
| *dagada* | 150 (31) | 172 (30) | 177 (32) | **49.1/.361** |
| *dada* | 166 (36) | 179 (38) | 208 (47) | **57.8/.398** |
| Mean duration (ms) | Females | | | |
| | no accent | H* preceding | L* preceding | s.s. $F(2,175)/\eta^2$ |
| *dagada* | 154 (24) | 175 (35) | 185 (34) | **81.6/.480** |
| *dada* | 177 (36) | 196 (50) | 201 (40) | **16.8/.163** |

alignment effects it has been invoked to account for, and potential future steps toward a more comprehensive theory of how adjacent tones influence each other.

### 5.4.1 Overall results

In this section, we briefly review the significance of our findings with respect to the acoustic correlates of $F_0$, energy and duration. Our analysis of $F_0$ correlates tested the tonal crowding hypothesis, which predicts that the $F_0$ peak for a H* pitch accent will be located in approximately the same place with respect to the accented syllable in all of our conditions, except the one in which the H* occurs on the final syllable of the phrase; in this case the peak is predicted to occur earlier, because of the crowding effect from the phrase accent and boundary tone which must be realized later in the same word. This prediction was supported

by the results: the peak location was not significantly different across the two target words with different numbers of syllables (*dagada* and *dada*), across the two non-final locations (early and late), across genders and for 17/20 individual speakers. In contrast, the peak was located significantly earlier in the accented syllable when the accent occurred on the phrase-final word (i.e. in the boundary condition), as predicted by the crowding hypothesis. Thus the results for H* support our hypothesis that in these utterances tonal crowding on the target word tends to shift the H* $F_0$ peak to occur earlier, presumably to allow room for the boundary tone to be realized.

The height of the $F_0$ peak for H* accents was also reduced for target words that occurred late in the phrase (i.e. in the *bnd* and *late-no-bnd* conditions, compared to the early condition). This is to be expected if these utterances were produced with overall global declination in $F_0$. An additional lowering of the $F_0$ peak was found for the *bnd* condition over the *late-no-bnd* condition, suggesting a possible truncation of the $F_0$ rise by the following L–L% tone combination. Note that truncation can apply to both the time domain and frequency domain.

Interestingly, the $F_0$ troughs associated with L* accents did not follow this pattern of proportionally earlier location in the syllable in response to tonal crowding; no significant differences were found between the *bnd* and the two *no-bnd* conditions. This raises the possibility that L*s are more variably realized, or perhaps are governed by a different set of principles than H* accents. We note that [AG07] found that the $F_0$ trough for the L in their L+H* accents was aligned more consistently than the $F_0$ peak for the H*, which they also view as evidence that L and H tones have different properties. Similarly, [ALM06] found significant differences in the behavior of the H and L targets in their study of Greek Polar Questions (see Section 5.4.3 below for further discussion).

Our predictions for comparisons of energy levels in accented syllables were that, as in earlier work, accented syllables would have higher energy levels, but that the difference might be less in the *bnd* condition because of falling subglottal pressure in the phrase-final word. The results were consistent with this prediction: energy differences between accented and unaccented syllables were smaller on average for the *bnd* case than for the other cases, regardless of gender and pitch accent. The energy difference between accented and unaccented syllables is less on average for L* than for H*, and less for L–L% compared to H–H%. This result is somewhat surprising because our analysis method, which takes a window of three pitch periods, was designed to neutralize any effects of $F_0$ levels themselves by using pitch-synchronous energy. Thus the smaller energy difference for L* and for L–L%, like the results for $F_0$ alignment, raises the possibility that L targets are governed by different principles than those that govern H targets.

Our results for duration showed that phrase-final syllable durations were, on average, longer when the immediately preceding syllable carried a pitch accent. This may be because the duration increase associated with a pitch accent can extend into the following syllable, as reported by [TW99]. Furthermore, it was found that, on average, syllables with preceding L* accents have longer duration than syllables with preceding H* accents. The longer duration attributed to syllables with preceding L* accents may be due to the shorter time required to achieve the preceding falling pitch. This durational difference in high and low pitch changes has been reported in [OE73] and [Sun79], and was later confirmed in [XS02]. A shorter time required to complete the preceding L* accent leaves more time for the final syllable. Similarly the longer time required to achieve a rising pitch can be attributed to the subsequent shorter duration of a final syllable which has a preceding H* accent. This result is yet one more indication that L* and H* pitch accents are not implemented in the same way. Speaker-

specific differences were also found, with some speakers not conforming to the general trends of longer duration for L* vs. H* and shortest duration for the case without a preceding pitch accented syllable.

Our duration results are in line with [BE94], [Kla76b], [TS07], and [TW99], among others, who found lengthening of the phrase-final syllable. In addition to lengthening of the phrase-final syllable, we found boundary-related lengthening in the main-stress (penultimate) syllable of both *dada* and *dagada* even when these words were unaccented. This finding extends the results for main-stress-syllable lengthening in phrase-final position utterance-medially ([TS07]) to main-stress-syllable lengthening in utterance-final position.

Overall, our results provide evidence that the acoustic correlates of high and low pitch accents in American English (for read speech) include $F_0$, energy and duration. By analyzing results separately for H* and L*, we have added to the growing evidence that these two kinds of pitch accents do not behave in precisely the same ways. In particular, the lack of evidence for effects of tonal crowding for L*s highlights the fact that the parameters of such crowding have not been thoroughly explored. We discuss some of the requirements for a full theory of tonal interaction in the final section of this chapter. Before turning to that discussion, however, we examine the significance of the varying results for individual speakers.

### 5.4.2 Individual speaker analyses

The results reported above have the advantage of being normalized for individual speakers, potentially increasing the power of the analyses. In addition, the availability of 20 sets of individual results gives some estimate of the range of variation in the acoustic correlates of tonal targets across speakers, and of the

differences in response to contextual factors such as tonal crowding. These differences are as important as the averaged trends, because they show the difficulty of placing some speakers into generalized models. The $F_0$ results, shown in Section 5.3.1, confirmed the hypothesis that the $F_0$ peak of an H* accented syllable would shift to an earlier point if there was another prosodic target (in this case, the boundary-related tones) which needed to be realized on the same word. This trend was observed for 8/10 males and 9/10 females. In this section, the results for the 2 male speakers, denoted by M1 and M2, and the 1 female speaker, denoted by F9, who did not conform to the general trends are discussed.

Figure 5.12 shows the scatter plot of the relative $F_0$ peak heights and their relative positions for the H* accented target word *dagada* for speaker M1. It can be seen that the boundary cases have minimal effect on the peak positions, which appear to be clustered around 75% of the normalized vowel duration. A similar plot was observed for the target word *dada*, with the peak positions appearing around 70%. While speaker M1 did not display any reliable shifts in peak position for any of the three cases, the height of the peak was consistent with general trends; that is, the *bnd* case had a smaller change in height than the *no-bnd-early(-daily)* and *no-bnd-late-daily* cases. Thus it appears that this speaker shares some of the more general response to tonal crowding, i.e. a smaller $F_0$ excursion for the H*, but not others, i.e. an earlier $F_0$ peak.

Speaker F9 had $F_0$ peak shifts aligned with the general trend for the target word *dagada*, but no peak movement could be seen for target word *dada*. The opposite was found for the heights of the $F_0$ peaks, with the target word *dagada* not conforming to the expected trend of having a lower relative height for the *bnd* condition; for this speaker, the $F_0$ peak height was on average about 30% higher for the *bnd* case than for the other cases (rather than lower, as was the

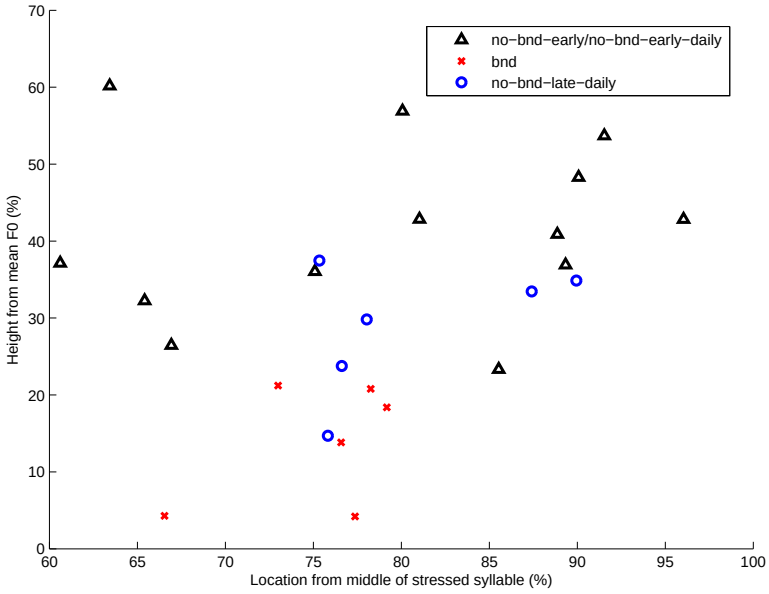



**Figure 5.12:** Scatter plot for the target word *dagada* showing relative $F_0$ peaks for H* and their relative positions in the accented target vowel for the male speaker M1 in three different contexts: 1. *no-bnd-early*/*no-bnd-early-daily* (triangles); 2. *bnd* (crosses); 3. *no-bnd-late-daily* (circles).

general trend). Again, this speaker showed some of the general responses to tonal crowding, but not as consistently as other speakers, and in at least one measure showed an idiosyncratic response.

Speaker M2 was unlike the majority of the speakers who had clearly detectable $F_0$ peaks for the H* accented target words, in that he had some $F_0$ contours whose shape was more similar to a down-stepped (!H*) accent. This occurred in 5/20 and 13/20 utterances for the target words *dagada* and *dada* respectively. Where there was a detectable $F_0$ peak, this speaker showed no shift of the peak to an earlier location under conditions of crowding for the *dagada* targets, and the relative height of the peak for the *bnd* case was higher than for the other cases.

Interestingly, informal listening showed that very little audible difference could be perceived between these three speakers (M1, M2 and F9) and the others. Table 5.10 summarizes the inconsistencies of these speakers when compared with the general trends for the *bnd* case, which were: 1) less $F_0$ peak shift, and 2) smaller $F_0$ peak change.

**Table 5.10:** Comparison of the speakers M1, M2 and F9's $F_0$ peak position and relative height consistencies with the general trends for the *bnd* case; a 'Yes'/'No' denotes agreement/disagreement while 'N/A' means no enough data was available.

| | ***dagada*** | | ***dada*** | |
|---|---|---|---|---|
| **Speaker** | Less $F_0$ peak shift for *bnd* case | Smaller $F_0$ peak change for *bnd* case | Less $F_0$ peak shift for *bnd* case | Smaller $F_0$ peak change for *bnd* case |
| M1 | No | Yes | No | Yes |
| M2 | No | No | N/A | N/A |
| F9 | Yes | No | No | Yes |

On average, the energy measure showed that for cases where the target word was located at the boundary (*bnd*), the normalized energy was lowest, followed by the *no-bnd-late-daily* cases, with the *no-bnd-early(-daily)* cases having the highest energy. This trend was generally adhered to by all of the speakers in the corpus, although for some speakers there was considerable overlap between the *bnd* and the *no-bnd-late-daily* cases; for example, speakers M3, M6 and M9 shown in Figure 5.5 for the target word *dagada*. Interestingly, these speakers also displayed the same behavior for the target word *dada*, suggesting perhaps that the energy measure is relatively consistent across the target words for each speaker.

Duration measure patterns were also fairly consistent across the target words for individual speakers. It was hypothesized that the phrase-final syllable would have extra lengthening if it was preceded by a pitch accent, regardless of the type of pitch accent. This hypothesis was shown to be true for all of the speakers if the $H^*$ and $L^*$ accents were considered together. However, when the pitch accent types were considered separately, as shown in Figure 5.11 for female speakers for the target word *dada*, some speakers (F5 and F8) had a shorter duration for the $H^*$ preceding case than the no-accent case. These two speakers also showed the same characteristics for the target word *dagada*. Other differences for other speakers were also found to be consistent across target words, suggesting that they are not random variation but controlled choices for parameter values.

While our generalized results provide a very compact way to describe the effects of tonal crowding, the individual differences described in this section show that acoustic correlates of tonal targets can vary substantially between speakers. It would be of interest to explore the possibility that these variations reflect different decisions about which cues to produce, vs. differences in degree of

control.

### 5.4.3 Theories of tonal crowding

While the effects of tonal crowding have been described in a number of contexts ([ALM06, AG07, GPN00, Od05]), the principles that govern these effects have not been fully and systematically explored. Moreover, tonal crowding effects can be seen as part of the more general question of how $F_0$ and other acoustic exponents of tonal targets are realized with respect to the segmental content of an utterance. A number of issues are raised by earlier work in this area, including the following:

1. *What are the options for a speaker when confronted by the need to realize several tonal targets in quick succession?* Several different mechanisms have been proposed, including truncation (i.e. a smaller $F_0$ movement, [GPN00], compression (i.e. a faster $F_0$ movement, [FJ98]), and deletion (i.e. elimination of one of the tonal targets, [Lev] for Turkish; [FJ98] for French). Another possible response is to move one or both of the two target realizations within their syllables so they occur further apart; articulating an $F_0$ movement earlier in its syllable would be one response of this type. Yet another possibility is to lengthen the segmental material, particularly the syllabic nucleus, i.e. slowing the speaking rate (at least temporarily) to make time for the realization of complex tone sequences. It appears that most of the speakers in this experiment, when confronted by tonal crowding from a pitch accent and boundary-related tones on the final two syllables of an utterance, chose a combination of a less-extreme $F_0$ movement, a longer duration of the syllable and earlier realization of the $F_0$ peak. The choices that speakers make among these possible alignment ad-

justment mechanisms, and the details of how those choices are realized, are in need of further investigation. For example, when two targets move apart in response to crowding, does just one of them move to an earlier location, or do both tonal targets move away from each other?

2. *What is the definition of crowding?* That is, how close to each other do two tonal targets have to be, in order to influence each other's realization, and how is this distance measured? Is it in terms of time increments, i.e. milliseconds? In terms of the number of voiced phonological segments that are available to carry the $F_0$? In terms of constituent structure (e.g. syllables)? Or in some combination of these scales? It has been suggested that, in order to eliminate the effect of one tonal target on another, it may be necessary to have two unstressed syllables between the two targeted syllables; others have suggested that the preferred target-syllable relationship is one target per syllable ([ALM06, Lev]). This decision is important because its result will inform the estimation of the preferred, non-crowded realization of a tonal target sequence, as well as the generation of appropriate algorithms for natural-sounding synthesis.

3. *Do different tonal target types respond differently to crowding?* That is, do all types of tonal targets follow the same principles of interaction? For example, do two adjacent pitch accents interact in the same way as a pitch accent followed by a boundary tone? Do bi-tonal targets behave differently from single tonal targets. High targets differently from Lows? Our results suggest that $L^*$ targets behave differently in response to tonal crowding than $H^*$ targets, and [SP90] results suggest that pre-nuclear and nuclear accents respond in largely similar ways. But the full scope of interactions among various tonal target types has not been investigated.

4. *Do different languages and dialects exhibit different principles of interaction?* For example, [GPN00] reported substantial differences in tonal target interactions in various dialects of British English. Similarly, Mücke and her colleagues report differences between Viennese and Dusseldorf German for resolution of such alignment issues ([MGB09]), and [Lev] reports target deletion by Turkish speakers to prevent tonal crowding.

While we don't yet have a clear picture of how these issues are resolved in American English speech, we can test the hypothesis that $F_0$ peaks seem to occur earlier in syllables that are lengthened phrase-finally, i.e. if the peak occurs at a fixed number of milliseconds from the V onset, it will seem to occur earlier in longer syllables. If final lengthening were responsible for the early $H^*$ peaks, then in our *bnd* condition the $H^*$ peak locations in the raw vowel durations should show one cluster of points for all three cases (early, late, boundary). Results are shown in Figure 5.13 which plots the time of the peak in milliseconds from the V onset against the height of the peak above the mean $F_0$, for two typical speakers, one male for target word *dagada* and one female for target word *dada*. For most speakers the peak position was significantly earlier for the *bnd* case, as per our results for normalized vowel durations, showing that the peak occurs earlier in absolute as well as relative terms under tonal crowding.

As [SP90] point out, the question of how tonal crowding affects the phonetic realization of pitch accents and other tonal targets is part of the larger general issue of tonal alignment. With this in mind, they evaluated several alternative mechanisms for differences in $F_0$ peak alignment across stimulus types, including *invariant duration* of the $F_0$ rise; *gestural overlap* and trunction; *tonal repulsion* with earlier gesture beginning; *phonological mediation* by the addition of extra beats to the metrical grid, lengthening the syllable with the result that the $F_0$

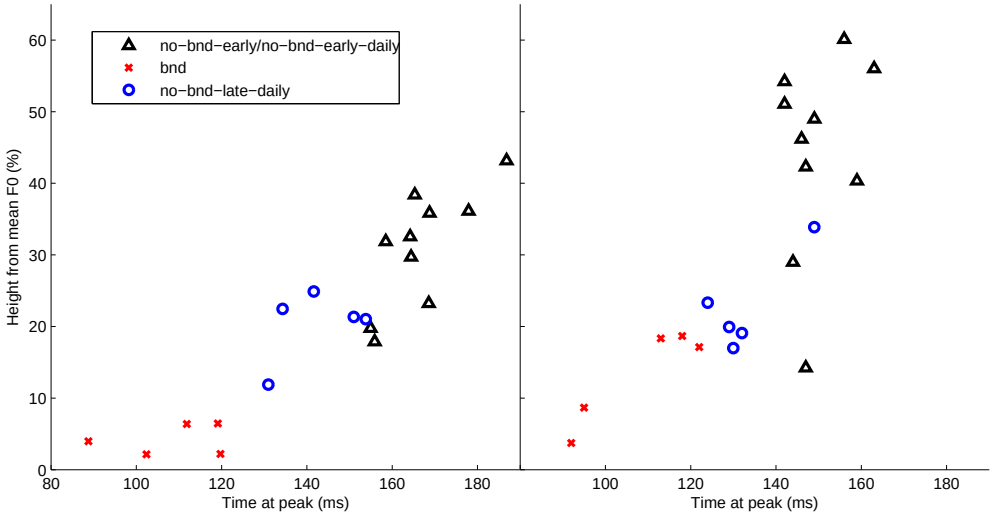



**Figure 5.13:** Scatter plot showing the times for the raw peak position in milliseconds from the V onset and the normalized peak heights for the three cases; *no-bnd-early(-daily)* (triangles), *bnd* (crosses) and *no-bnd-late-daily* (circles). The left panel shows the times and heights for a typical male speaker for the target word *dagada* and the right panel shows the results for a typical female speaker for the target word *dada*.

peak occurs earlier; and *sonority profile*.

In summary, although a clear picture of the factors that govern tonal alignment for different tonal target types in various contexts and in different languages is still emerging, studies such as this one have begun to reveal some aspects of these patterns for individual languages. It is clear that considerable further research is needed to clarify the factors that govern the alignment of $F_0$ contours with the words and syllables of a spoken utterance, and how the alignment changes under conditions of tonal crowding.

## 5.5 Conclusion

This study compares the acoustic characteristics of two types of pitch accents in American English ($H^*$ and $L^*$) in three types of locations within the phrase (in early and late non-phrase-final words and phrase-final words), in two types of target words (two- and three-syllable words) with penultimate lexical stress. Results for $F_0$ show that for most speakers and for both target words, the $F_0$ peak for a nuclear $H^*$ accent occurs earlier in conditions of tonal crowding due to phrase-final boundary tones in the same word, and is realized with a lower $F_0$. In contrast, $F_0$ troughs for $L^*$ accents did not show the same effects of tonal crowding, suggesting that $H^*$ and $L^*$ accents may not be realized according to the same principles. Comparison across 20 individual speakers (10 male and 10 female) revealed that the general findings are robust, but that 3 speakers did not comport with all aspects of the general findings, raising the possibility that some speakers may employ idiosyncratic cue patterns. Analysis also showed that energy levels decrease across the utterance, and that a phrase-final syllable is longer if the immediately-preceding syllable in the final word is accented than if it is not. Taken together, these results highlight the importance of taking context

into account for prosodic analysis.

# CHAPTER 6

# Summary and Future Work

## 6.1 Summary

In this dissertation, the analysis and properties of the voice source with respect to voice quality were presented.

Chapter 1 introduced the background information on human speech production and the linear speech production model including the voice source and vocal tract components. Also presented were some of the existing source estimation methods and the definitions and terminology used in voice quality and prosody analysis.

In Chapter 2, a new source model was derived from glottal area waveform obtained via high-speed imaging. These direct observations of the vocal folds revealed some characteristics of the source which were not possible to represent with existing source models. A new source model was proposed which better captured the observed glottal area waveforms.

Chapter 3 presented a different approach to the traditional inverse-filtering technique in estimating the source signal from speech data. Using the LF and the new proposed source model, a novel codebook search approach was used for source estimation.

In Chapter 4, a new software, `VoiceSauce`, was introduced which simplified

the calculation of common voice source related measures such as $H_1^*-H_2^*$, $H_1^*-A_3^*$, *CPP* and *HNR*, to name a few. This software was used in three applications, voice quality analysis, automatic gender classification, and prosody analysis.

Chapter 5 analyzed $F_0$, energy and duration in relation to intonational pitch accents. The effects of having multiple pitch accents in close proximity to each other were also examined.

The following three sections summarize the key results of the analysis and properties of the voice source with respect to voice quality.

### 6.1.1 Source modeling and estimation

In this study, direct observations of the source were made by filming the vocal folds through high-speed imaging. From these observations it was found that existing source models were deficient in two main aspects: (1) the duration of the opening phase has often been assumed to be longer than the duration of the closing phase, but the reverse was observed, and (2) the speed of opening and closing was observed to be much faster than what could be specified in existing models. A new four-parameter ($OQ$, $\alpha$, $S_{op}$ and $S_{cp}$) source model, derived from the popular LF model, was proposed to rectify these two aspects. Results showed that the proposed new source model provided a better fit for the glottal area waveforms obtained from the high-speed imaging than the LF model.

A new source estimation technique, utilizing a codebook approach with spectral analysis-by-synthesis, was introduced which effectively inverse-filtered the speech signal, with the source signal instead of the vocal tract as in traditional inverse-filtering schemes. Results comparing the fitted model parameters with the estimated model parameters showed that while there were good correlations for the parameter $OQ$ and moderate correlations for the parameter $S_{op}$, there were

also no significant correlations for the parameters $\alpha$ and $S_{cp}$. It was hypothesized that this may have been due to the lack of influence these two parameters had on the spectral harmonic magnitudes, which were used for the analysis-by-synthesis in the frequency domain. Analysis showed that high $F_0$ phonations had the most estimation errors due to the difficulty with estimating the formant frequencies for high-pitched voices. Error analyses showed that it was important to use an accurate source model and reasonable formant frequency constraints to obtain good source signal estimates.

### 6.1.2 Correlates of voice quality

The `VoiceSauce` application, which simplified the process of calculating voice source related measures was used in three different scenarios: voice quality analysis, automatic gender classification and prosody analysis.

Using the `VoiceSauce` application, the voice source related measures $H_1-H_2$, $H_2-H_4$, $H_1-A_1$, $H_1-A_2$, $H_1-A_3$, *Energy*, *CPP*, and *HNR* were calculated using the audio data which was collected synchronously with the high-speed imaging. By comparing these measures with the fitted model parameters, it was found that on average, the open quotient parameter $OQ$ and the spectral tilt measures, $H_1 - A_2$ and $H_1 - A_3$, were affected by both voice quality and glottal gaps. The asymmetry parameter $\alpha$, and the measures *CPP* and the three *HNR* measures were found to be predominately affected by voice quality, especially of the breathy type. This indicated the presence of more spectral noise for breathy phonations and suggested that asymmetry may be an important part of how the noise is generated. An interesting positive correlation was found between the parameters $OQ$ and $S_{op}$, which was not replicated for the parameter $S_{cp}$.

Automatic gender classification using speech measures was found to improve

with the addition of voice-source related measures $H_1^* - H_2^*$ and $H_1^* - A_3^*$. The improvement was particular pronounced for the age groups 10–11 and 12–13 year olds where traditional features, such as $F_0$ and formant frequencies, were less useful.

In the analysis of prosody, pitch accents were, as expected, found to be clearly marked by differences in $F_0$ contours. However, there was a difference in the positions of the $F_0$ maxima for high ($H^*$) accented syllables and the $F_0$ minima for low ($L^*$) accented syllables, in that the $F_0$ maxima were often found towards the end of the accented syllable. Lexical stress was found to be denoted by decreases in $H_1^* - H_2^*$, indicating a tenser voice on the stressed syllable. The spectral tilt measure $H_1^* - A_3^*$ was seen to be gender dependent, but for boundary-related tones, the measure generally decreased towards the end of the word, denoting an increase in high-frequency energy.

### 6.1.3 Correlates of pitch accents

Through the use of a carefully designed speech corpus, analysis of the $F_0$ contours revealed that, in American English, the position of the $F_0$ peak for a high accent occurs earlier in conditions of tonal crowding due to phrase-final boundary tones in the same word, and is realized with a lower $F_0$. In contrast, $F_0$ troughs for low accented syllables did not show the same effects of tonal crowding, suggesting that high and low accents may be realized with different principles. Analysis of the energy levels showed that on average, energy measures were greater if the target word was at the start of a sentence than at the end. However, the measures were seen to be dependent on the type of accent, position of the accent and the proximity of the boundary tone. Duration analysis found that a phrase-final syllable is longer if the immediately-preceding syllable in the final word is

accented than if it is not. Overall, the results emphasized the importance of taking context into account when performing prosodic analysis.

## 6.2 Unsolved issues and outlook

This dissertation examined the voice source with respect to voice quality from two different perspectives. The first, through the direct observation of the vocal folds and the second, through analysis of voice source related measures. Direct observations of the source allowed for a more accurate source model to be created, however, there are still some unsolved issues. In the glottal area waveforms shown in Figure 2.4 and Figures A.1–A.5, it could be seen (more apparent for subject FM3) that some of the waveforms had a "shoulder" before the main peak. It is not clear what causes this shoulder effect; no voice source model, to date, is able to capture this type of waveform. The glottal area waveforms used in this work were not normalized for the distance of the camera to the vocal folds. Thus, it was not possible to compare certain characteristics between phonations, such as the maximum or minimum areas. For example, it is conceivable that a pressed phonation should have a much smaller maximum glottal opening than a breathy phonation. Continuous recordings of subjects varying their voice qualities would help in this regard. The glottal gap and its related effects is another area which requires more research. Preliminary analysis on a small number of subjects showed that breathy phonations typically have incomplete glottal closures as well as increased spectral noise coupled with decreased source asymmetry (parameter $\alpha$). At present, it is not clear which is the cause and which is the effect; for example, is the act of producing spectral noise causing the asymmetry to decrease or does a decrease in asymmetry produce more spectral noise? More data with varying phonation types from more subjects are needed to address this issue.

The work in the first part of this dissertation provided a first look at the voice source signal, as represented by the glottal area waveform, for three voice qualities, pressed, normal and breathy. Future exploration should focus on continuous transitions between voice qualities; for example, from pressed to breathy, from low $F_0$ to high $F_0$ and from an /I/ vowel to an /æ/ vowel. Observations of these types of transitions would shed more light into the workings of the vocal folds. With improving technology, it may soon be possible to achieve the ultimate goal in voice source analysis, that of direct observations of the voice source in continuous, natural speech.

The second part of this dissertation focused on voice source/quality analysis through the use of voice source related measures. Through the high-speed imaging of the vocal folds, correlations could be made between the physiological data and the acoustic data. This provided many interesting results, including the asymmetry and harmonic-to-noise ratio data. More data from more subjects would allow the relationship between the physiological data and the acoustic data to be quantified mathematically, which would vastly improve the efficiency of voice source studies.

Solving the issues described here would lead to a deeper understanding of the voice source and its effects on voice quality. This knowledge can help improve practical applications such as speech analysis, speech coding, speaker identification, speech recognition and medical applications.

# APPENDIX A

## Averaged Glottal Area Waveforms

The averaged glottal area waveforms of subjects FM2–3 and M1–3 are shown in Figures A.1–A.5. The waveforms were created by the method described in Section 2.2.3. Note that subject FM2 was unable to produce a pressed phonation with normal $F_0$ while subject M1 was not able to produce phonations with low $F_0$ for any voice quality. The averaged glottal area waveforms for subject FM1 can be found in Figure 2.4.

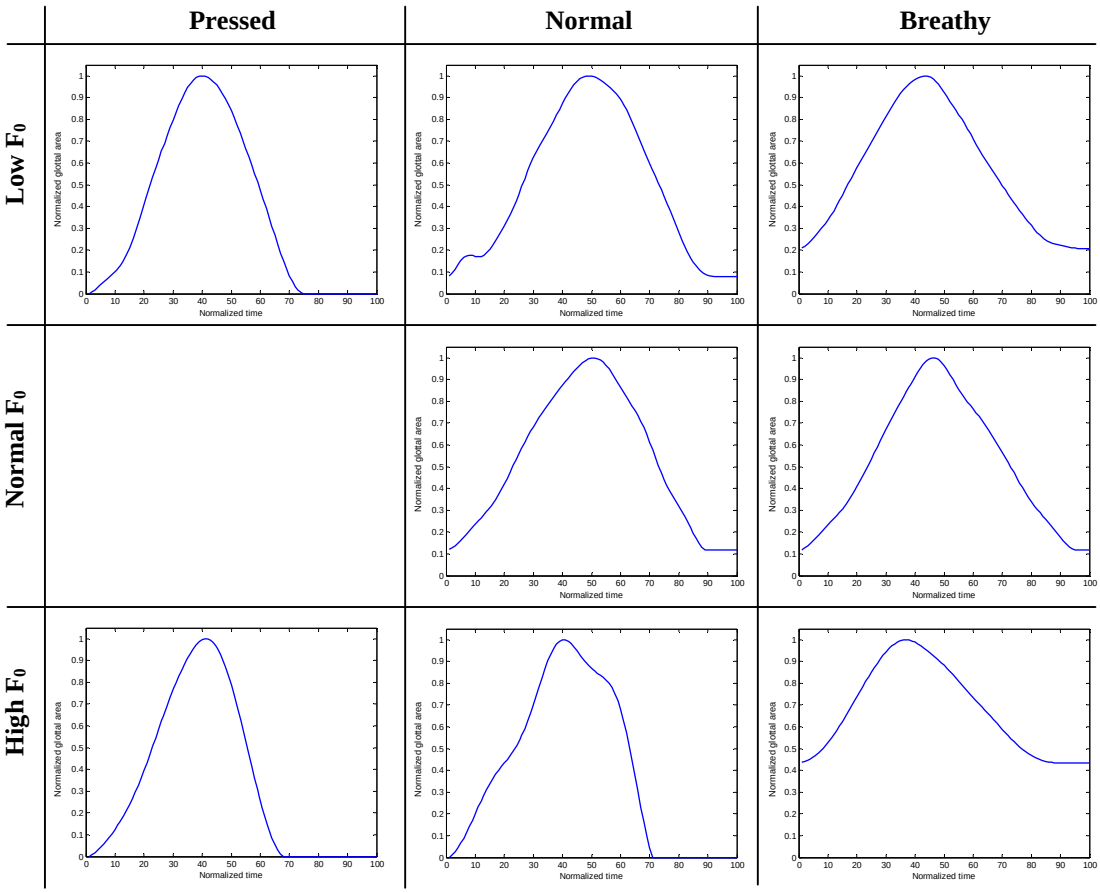



**Figure A.1:** The averaged glottal waveforms for the nine phonation combinations for subject FM2. $F_0$ (low, normal and high) was varied quasi-orthogonally with voice quality (pressed, normal and breathy). Data for the pressed phonation with normal $F_0$ was not available for this subject.

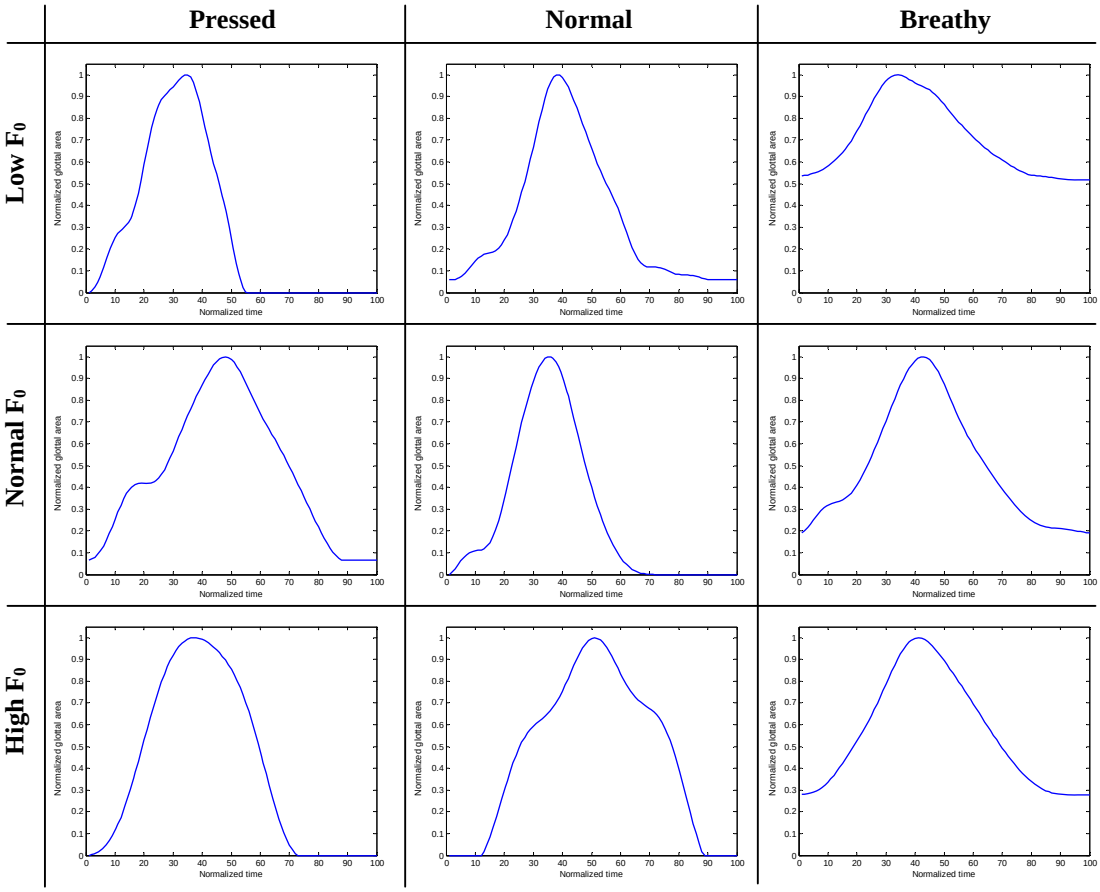



**Figure A.2:** The averaged glottal waveforms for the nine phonation combinations for subject FM3. $F_0$ (low, normal and high) was varied quasi-orthogonally with voice quality (pressed, normal and breathy).

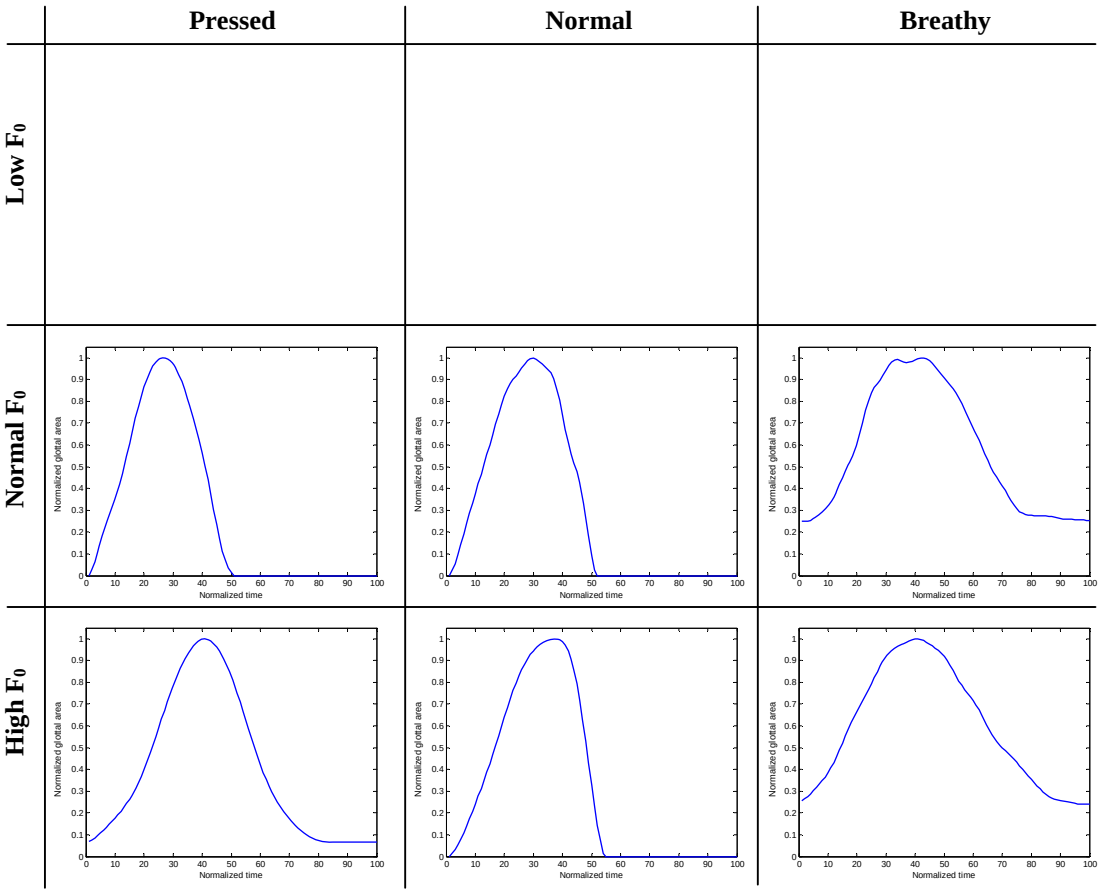



**Figure A.3:** The averaged glottal waveforms for the nine phonation combinations for subject M1. $F_0$ (low, normal and high) was varied quasi-orthogonally with voice quality (pressed, normal and breathy). Data for the low $F_0$ phonations was not available for this subject.

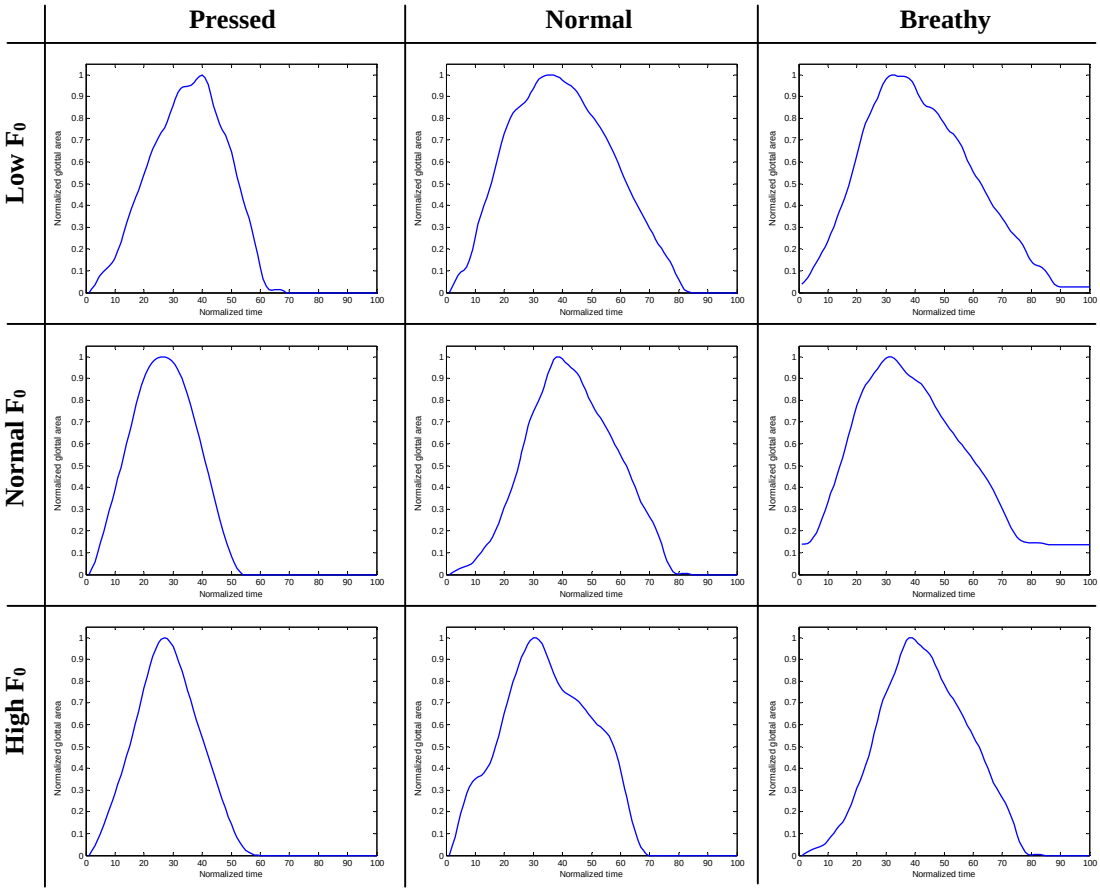



**Figure A.4:** The averaged glottal waveforms for the nine phonation combinations for subject M2. $F_0$ (low, normal and high) was varied quasi-orthogonally with voice quality (pressed, normal and breathy).

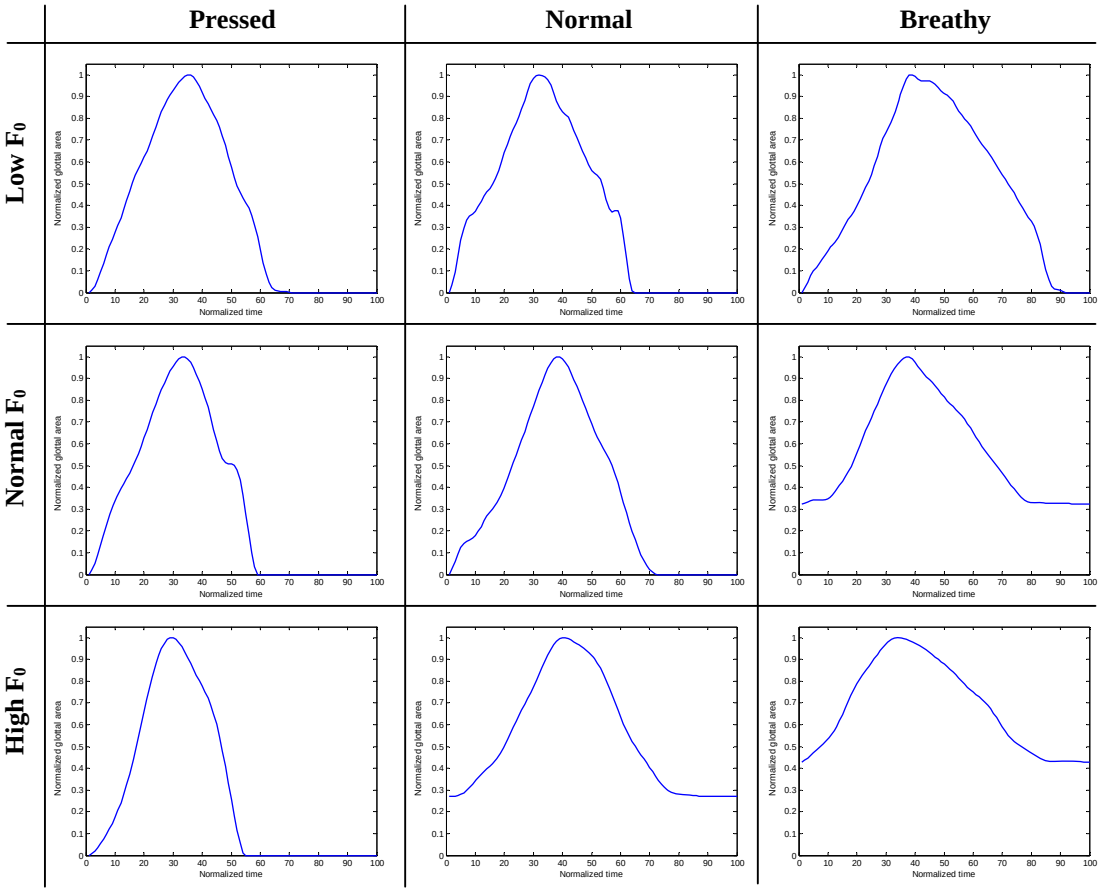



**Figure A.5:** The averaged glottal waveforms for the nine phonation combinations for subject M3. $F_0$ (low, normal and high) was varied quasi-orthogonally with voice quality (pressed, normal and breathy).

# APPENDIX B

# Glottal Area Model Fitting Performance of the Proposed New Source Model

Figures B.1–B.6 show the model fitting performance of the proposed new source model for the averaged glottal area waveforms described in Section 2.2.3. Tables B.1–B.6 lists the voice source parameters from the model fitting. Note that subject FM2 was unable to produce a pressed phonation with normal $F_0$, while subject M1 was not able to produce phonations with low $F_0$ for any voice quality.

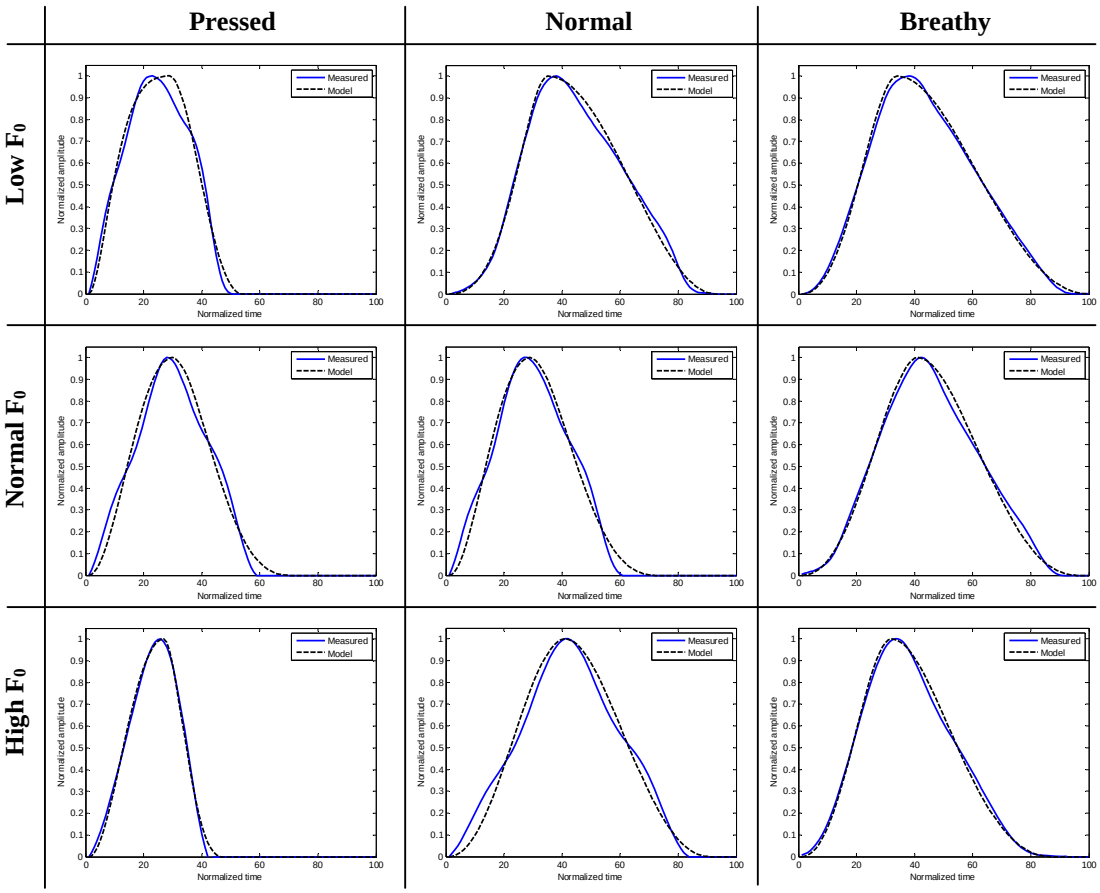



**Figure B.1:** Model fitting performance of the proposed new source model for subject FM1.

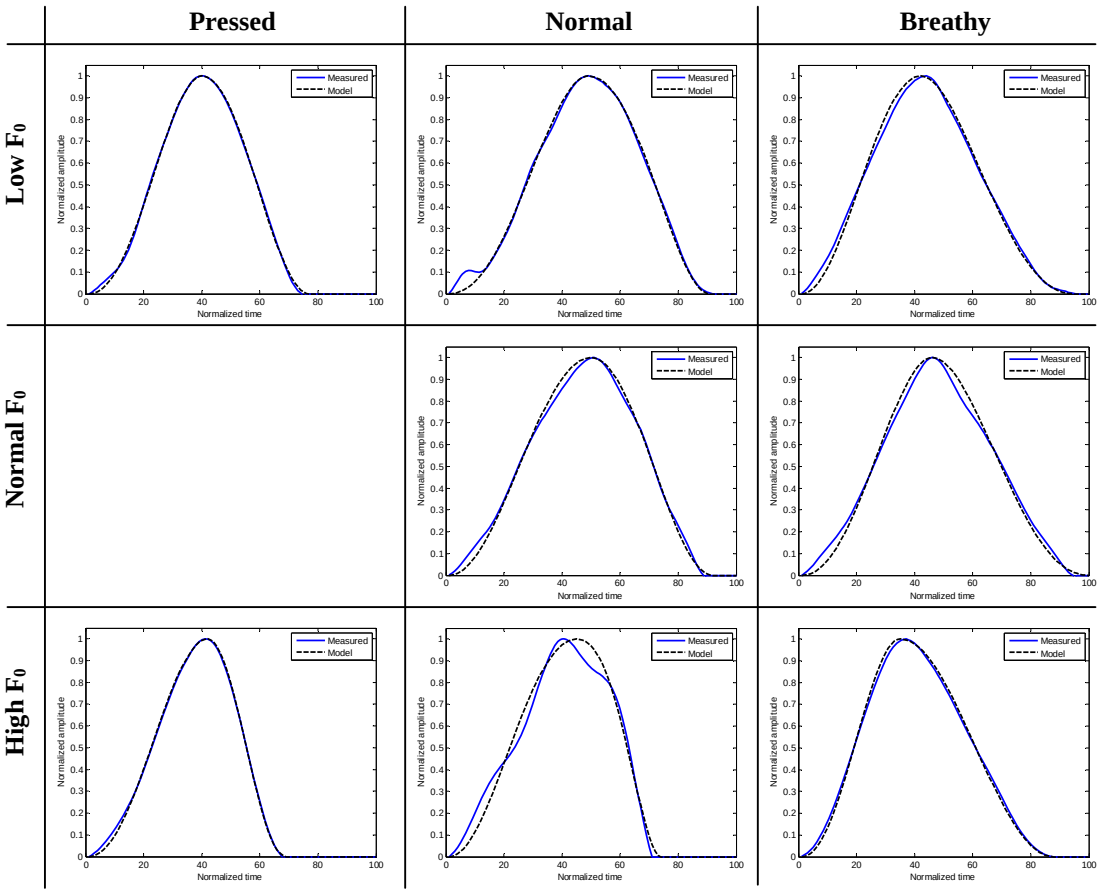



**Figure B.2:** Model fitting performance of the proposed new source model for subject FM2.

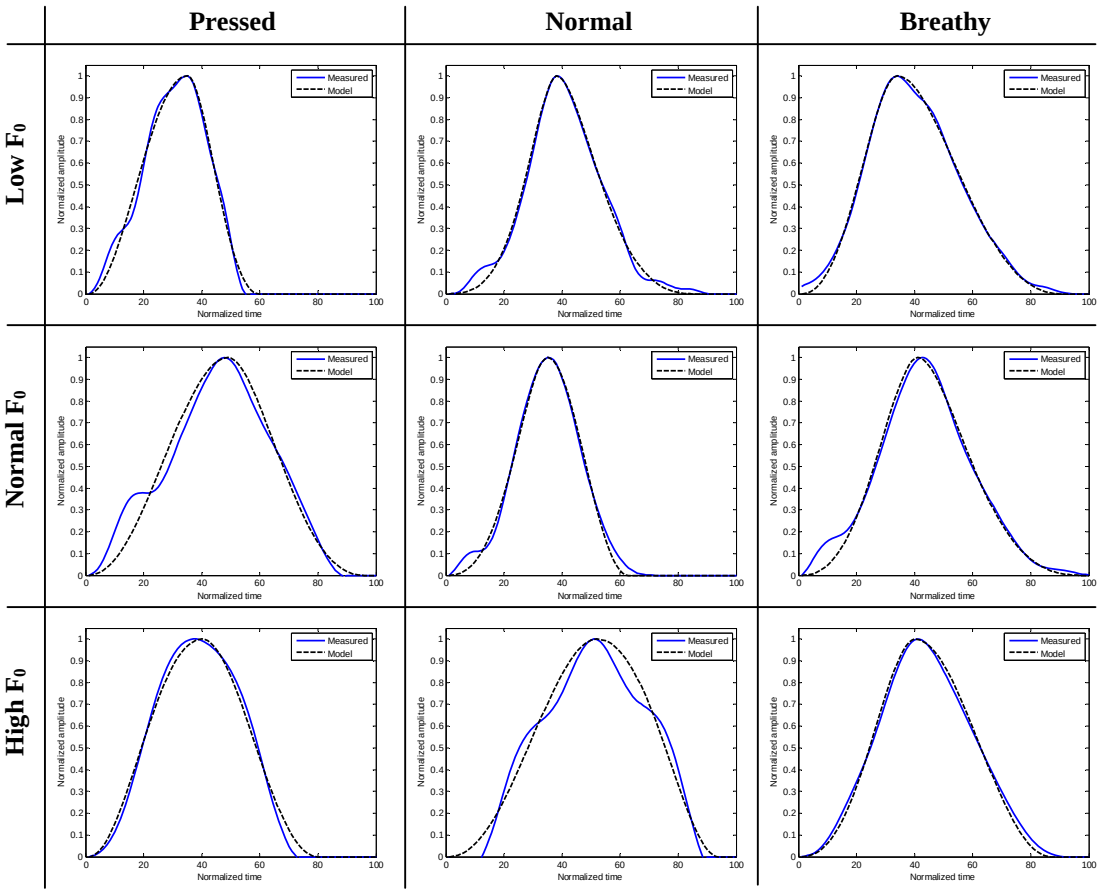



**Figure B.3:** Model fitting performance of the proposed new source model for subject FM3.

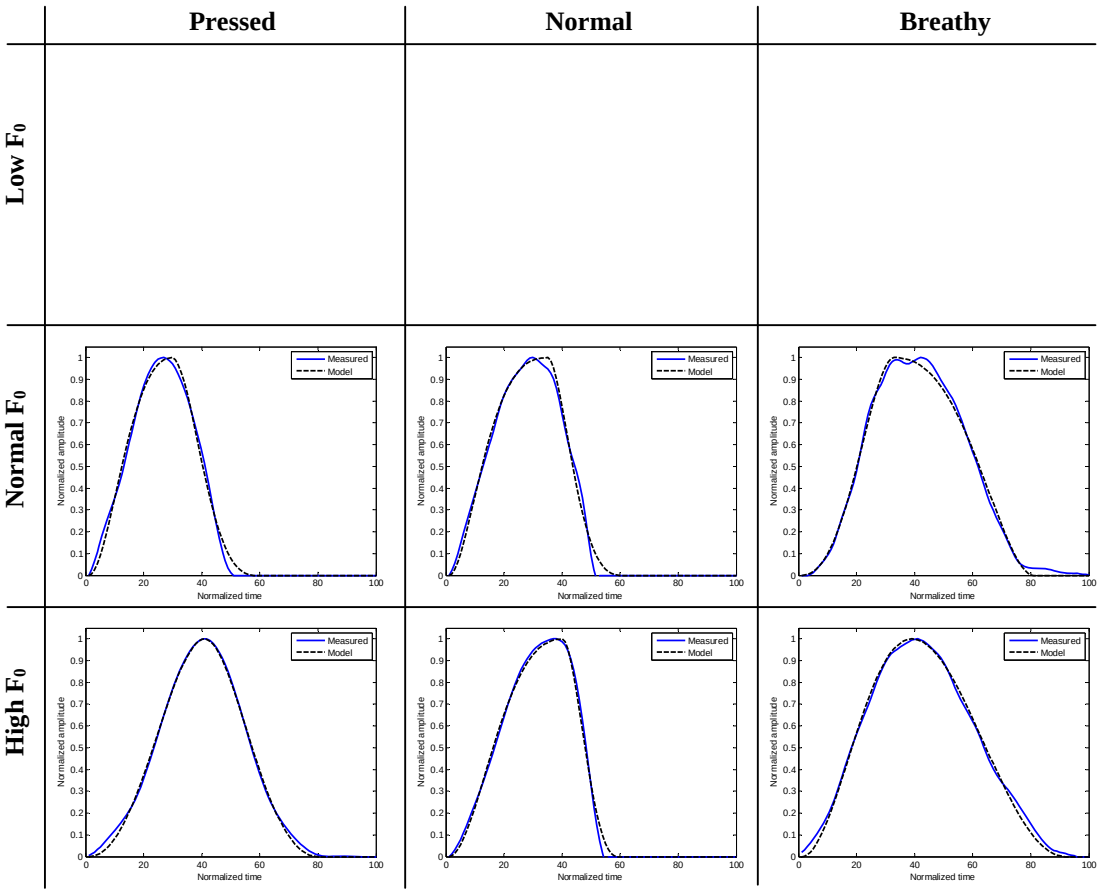



**Figure B.4:** Model fitting performance of the proposed new source model for subject M1.

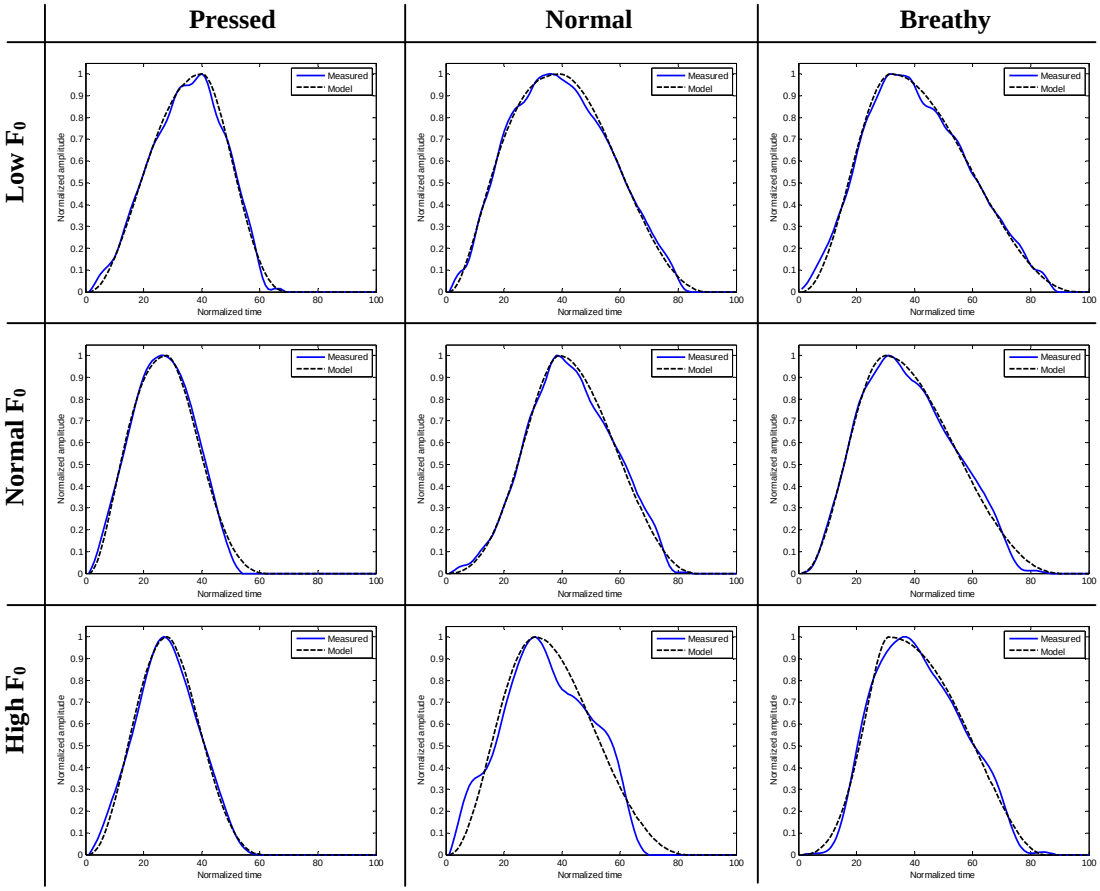



**Figure B.5:** Model fitting performance of the proposed new source model for subject M2.

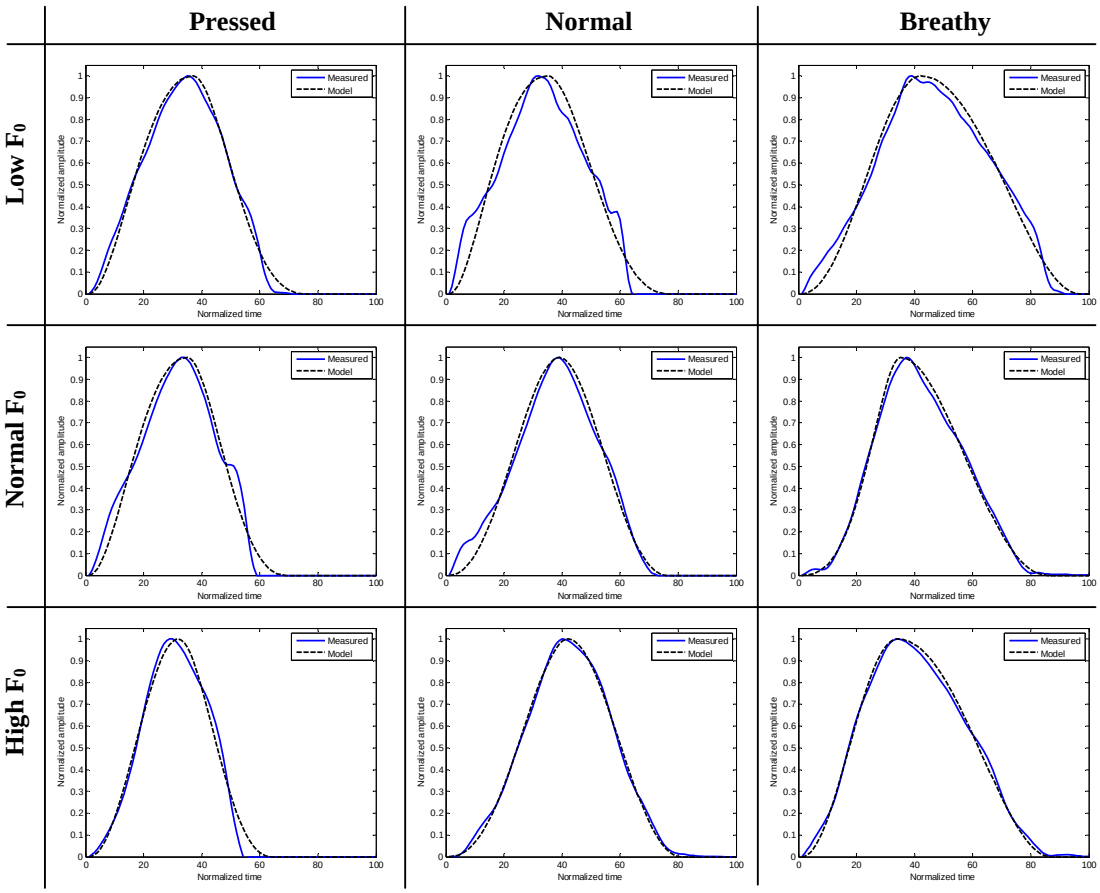



**Figure B.6:** Model fitting performance of the proposed new source model for subject M3.

**Table B.1:** Voice source parameters from the model fit (see Figure B.1) for subject FM1. "G. gap" denotes the existence/absence of the glottal gap.

| Param. | Pressed | | | Normal | | | Breathy | | |
|---|---|---|---|---|---|---|---|---|---|
| | low | norm. | high | low | norm. | high | low | norm. | high |
| $OQ$ | .4950 | .6512 | .4785 | .9518 | .6545 | .9595 | .9985 | .9996 | .9188 |
| $\alpha$ | .4710 | .4328 | .5298 | .3568 | .4163 | .4304 | .3451 | .3980 | .3428 |
| $S_{op}$ | .3661 | .4739 | .4637 | .6598 | .4651 | .5134 | .5635 | .5867 | .5536 |
| $S_{cp}$ | .3867 | .5708 | .6309 | .5252 | .5384 | .6253 | .5661 | .6113 | .6342 |
| G. gap | No | No | No | Yes | No | No | Yes | Yes | Yes |

**Table B.2:** Voice source parameters from the model fit (see Figure B.2) for subject FM2. "G. gap" denotes the existence/absence of the glottal gap. Pressed, normal $F_0$ phonations were not available for this subject.

| Param. | Pressed | | | Normal | | | Breathy | | |
|---|---|---|---|---|---|---|---|---|---|
| | low | norm. | high | low | norm. | high | low | norm. | high |
| $OQ$ | .7548 | – | .6942 | .9146 | .9410 | .7261 | .9718 | .9996 | .9078 |
| $\alpha$ | .5119 | – | .6006 | .5339 | .5396 | .6033 | .4403 | .4499 | .3921 |
| $S_{op}$ | .5529 | – | .5171 | .5467 | .4762 | .4805 | .4686 | .5480 | .5067 |
| $S_{cp}$ | .4686 | – | .5435 | .4817 | .5504 | .3834 | .6163 | .5686 | .5606 |
| G. gap | No | – | No | Yes | Yes | No | Yes | Yes | Yes |

**Table B.3:** Voice source parameters from the model fit (see Figure B.3) for subject FM3. "G. gap" denotes the existence/absence of the glottal gap.

| Param. | Pressed | | | Normal | | | Breathy | | |
|---|---|---|---|---|---|---|---|---|---|
| | low | norm. | high | low | norm. | high | low | norm. | high |
| $OQ$ | .5815 | .9786 | .7169 | .8654 | .6735 | .9499 | .9115 | .9955 | .9425 |
| $\alpha$ | .5821 | .4992 | .4742 | .4325 | .5026 | .5294 | .3619 | .4184 | .4273 |
| $S_{op}$ | .4883 | .5114 | .5454 | .6981 | .6402 | .5432 | .5878 | .6008 | .5822 |
| $S_{cp}$ | .5628 | .6359 | .3480 | .6885 | .6034 | .4609 | .6057 | .6974 | .6090 |
| G. gap | No | Yes | No | Yes | No | No | Yes | Yes | Yes |

**Table B.4:** Voice source parameters from the model fit (see Figure B.4) for subject M1. "G. gap" denotes the existence/absence of the glottal gap. Low $F_0$ phonations were not available for this subject.

| Param. | Pressed | | | Normal | | | Breathy | | |
|---|---|---|---|---|---|---|---|---|---|
| | low | norm. | high | low | norm. | high | low | norm. | high |
| $OQ$ | – | .5355 | .8288 | – | .6045 | .5655 | – | .7862 | .9865 |
| $\alpha$ | – | .5387 | .4922 | – | .5582 | .7012 | – | .4188 | .4030 |
| $S_{op}$ | – | .4013 | .5448 | – | .3380 | .3843 | – | .5810 | .4348 |
| $S_{cp}$ | – | .5836 | .6336 | – | .6668 | .5657 | – | .3786 | .6037 |
| G. gap | – | No | Yes | – | No | No | – | Yes | Yes |

**Table B.5:** Voice source parameters from the model fit (see Figure B.5) for subject M2. "G. gap" denotes the existence/absence of the glottal gap.

| Param. | Pressed | | | Normal | | | Breathy | | |
|---|---|---|---|---|---|---|---|---|---|
| | low | norm. | high | low | norm. | high | low | norm. | high |
| $OQ$ | .6675 | .5625 | .6290 | .8913 | .8663 | .7681 | .9801 | .8632 | .8719 |
| $\alpha$ | .5745 | .5067 | .4281 | .3935 | .4329 | .3841 | .3288 | .3296 | .3233 |
| $S_{op}$ | .4694 | .3758 | .5039 | .4043 | .6331 | .4936 | .4993 | .5150 | .4697 |
| $S_{cp}$ | .5421 | .5771 | .6436 | .5149 | .5423 | .5023 | .5646 | .5221 | .5526 |
| G. gap | No | No | No | No | No | No | Yes | Yes | Yes |

**Table B.6:** Voice source parameters from the model fit (see Figure B.6) for subject M3. "G. gap" denotes the existence/absence of the glottal gap.

| Param. | Pressed | | | Normal | | | Breathy | | |
|---|---|---|---|---|---|---|---|---|---|
| | low | norm. | high | low | norm. | high | low | norm. | high |
| $OQ$ | .6030 | .6587 | .5674 | .7455 | .7698 | .8203 | .9802 | .8722 | .9054 |
| $\alpha$ | .4020 | .5054 | .5630 | .4402 | .5010 | .5062 | .4222 | .3940 | .3702 |
| $S_{op}$ | .4916 | .4409 | .4969 | .4280 | .5442 | .5707 | .5187 | .6536 | .4895 |
| $S_{cp}$ | .5236 | .5566 | .4750 | .5672 | .5803 | .5584 | .4954 | .5588 | .5201 |
| G. gap | No | No | No | No | No | Yes | No | Yes | Yes |

# APPENDIX C

# Voice Source Estimation Results for each Subject

Tables C.1, C.2 and C.3 list the MSE values for the voice source estimation using the proposed new source model for the Snack-, manual- and constant-based formant frequency constraints respectively. Results are listed for individual subjects in terms of voice quality (pressed, normal and breathy) and $F_0$ type (low, normal and high). '–' denotes that data was not available for a particular phonation.

Figures C.1–C.6 show the measured and estimated glottal source waveforms for all six subjects. Note that the measured waveforms have the DC-offset removed. The estimated waveforms were from the Snack- and manual-based formant constraints.

**Table C.1:** MSE values for source estimation using Snack-based formant constraints with the proposed new source model; results listed in terms of voice quality (pressed, normal and breathy) and $F_0$ type (low, normal and high). '–' denotes data was not available for a particular phonation.

| **Subject** | $F_0$ **type** | **Voice quality** | | |
|---|---|---|---|---|
| | | Pressed | Normal | Breathy |
| FM1 | Low | 0.0035 | 0.0290 | 0.0277 |
| | Normal | 0.1964 | 0.0290 | 0.0123 |
| | High | 0.0205 | 0.0873 | 0.0276 |
| FM2 | Low | 0.0046 | 0.0032 | 0.0231 |
| | Normal | – | 0.0220 | 0.0250 |
| | High | 0.0382 | 0.2684 | 0.0095 |
| FM3 | Low | 0.0277 | 0.0262 | 0.0242 |
| | Normal | 0.0148 | 0.0954 | 0.0306 |
| | High | 0.2158 | 0.0495 | 0.2995 |
| M1 | Low | – | – | – |
| | Normal | .0036 | 0.0067 | 0.0555 |
| | High | 0.0405 | 0.1088 | 0.0313 |
| M2 | Low | 0.0339 | 0.0543 | 0.0517 |
| | Normal | 0.0027 | 0.0238 | 0.0146 |
| | High | 0.0064 | 0.0281 | 0.0226 |
| M3 | Low | 0.0092 | 0.0409 | 0.0039 |
| | Normal | 0.0141 | 0.0224 | 0.0336 |
| | High | 0.0303 | 0.0206 | 0.0385 |

**Table C.2:** MSE values for source estimation using manual-based formant constraints with the proposed new source model; results listed in terms of voice quality (pressed, normal and breathy) and $F_0$ type (low, normal and high). '–' denotes data was not available for a particular phonation.

| **Subject** | $F_0$ **type** | **Voice quality** | | |
|---|---|---|---|---|
| | | Pressed | Normal | Breathy |
| FM1 | Low | 0.0037 | 0.0290 | 0.0314 |
| | Normal | 0.1774 | 0.0161 | 0.0024 |
| | High | 0.0018 | 0.0248 | 0.0260 |
| FM2 | Low | 0.0026 | 0.0181 | 0.0231 |
| | Normal | – | 0.0163 | 0.0140 |
| | High | 0.0725 | 0.0510 | 0.0110 |
| FM3 | Low | 0.0277 | 0.0044 | 0.0062 |
| | Normal | 0.0116 | 0.0552 | 0.0225 |
| | High | 0.0089 | 0.2313 | 0.0190 |
| M1 | Low | – | – | – |
| | Normal | 0.0048 | 0.0040 | 0.0555 |
| | High | 0.0463 | 0.0302 | 0.0353 |
| M2 | Low | 0.0339 | 0.0543 | 0.0517 |
| | Normal | 0.0027 | 0.0210 | 0.0146 |
| | High | 0.0380 | 0.0246 | 0.0216 |
| M3 | Low | 0.0092 | 0.0409 | 0.0039 |
| | Normal | 0.0119 | 0.0140 | 0.0303 |
| | High | 0.0303 | 0.0206 | 0.0647 |

**Table C.3:** MSE values for source estimation using constant-based formant constraints with the proposed new source model; results listed in terms of voice quality (pressed, normal and breathy) and $F_0$ type (low, normal and high). '–' denotes data was not available for a particular phonation.

| **Subject** | $F_0$ **type** | **Voice quality** | | |
|---|---|---|---|---|
| | | Pressed | Normal | Breathy |
| FM1 | Low | 0.0030 | 0.0272 | 0.0277 |
| | Normal | 0.1912 | 0.0180 | 0.0200 |
| | High | 0.0062 | 0.0515 | 0.0308 |
| FM2 | Low | 0.0047 | 0.0142 | 0.0231 |
| | Normal | – | 0.0212 | 0.0175 |
| | High | 0.0693 | 0.0497 | 0.0080 |
| FM3 | Low | 0.0253 | 0.0209 | 0.0110 |
| | Normal | 0.0067 | 0.0762 | 0.0225 |
| | High | 0.0084 | 0.0426 | 0.0116 |
| M1 | Low | – | – | – |
| | Normal | 0.0048 | 0.0040 | 0.0555 |
| | High | 0.0463 | 0.0302 | 0.0353 |
| M2 | Low | 0.0339 | 0.0543 | 0.0517 |
| | Normal | 0.0027 | 0.0210 | 0.0146 |
| | High | 0.0062 | 0.0389 | 0.0245 |
| M3 | Low | 0.0092 | 0.0366 | 0.0039 |
| | Normal | 0.0117 | 0.0224 | 0.0295 |
| | High | 0.0303 | 0.0206 | 0.0647 |

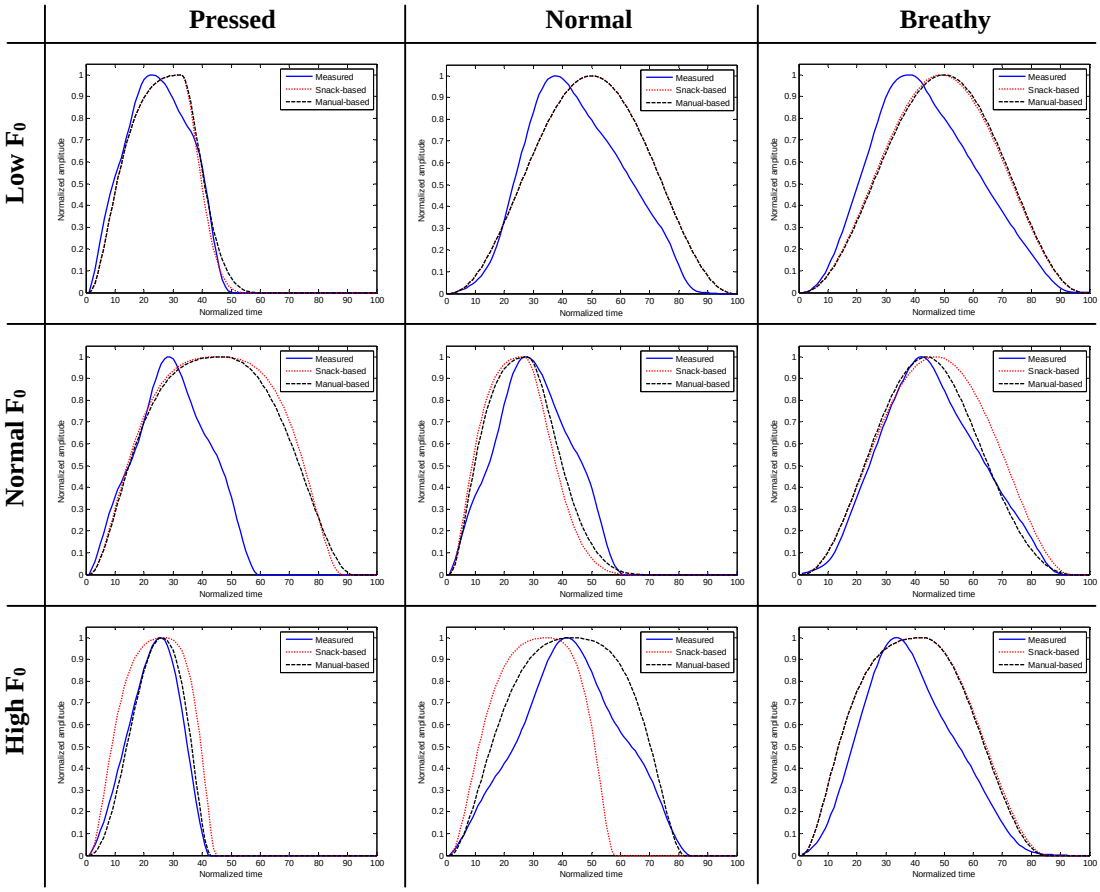



**Figure C.1:** Plot of the measured (solid line) and estimated glottal area waveforms for subject FM1. The estimated waveforms are from the Snack-based (dotted line) and manual-based (dashed line) formant constraints.

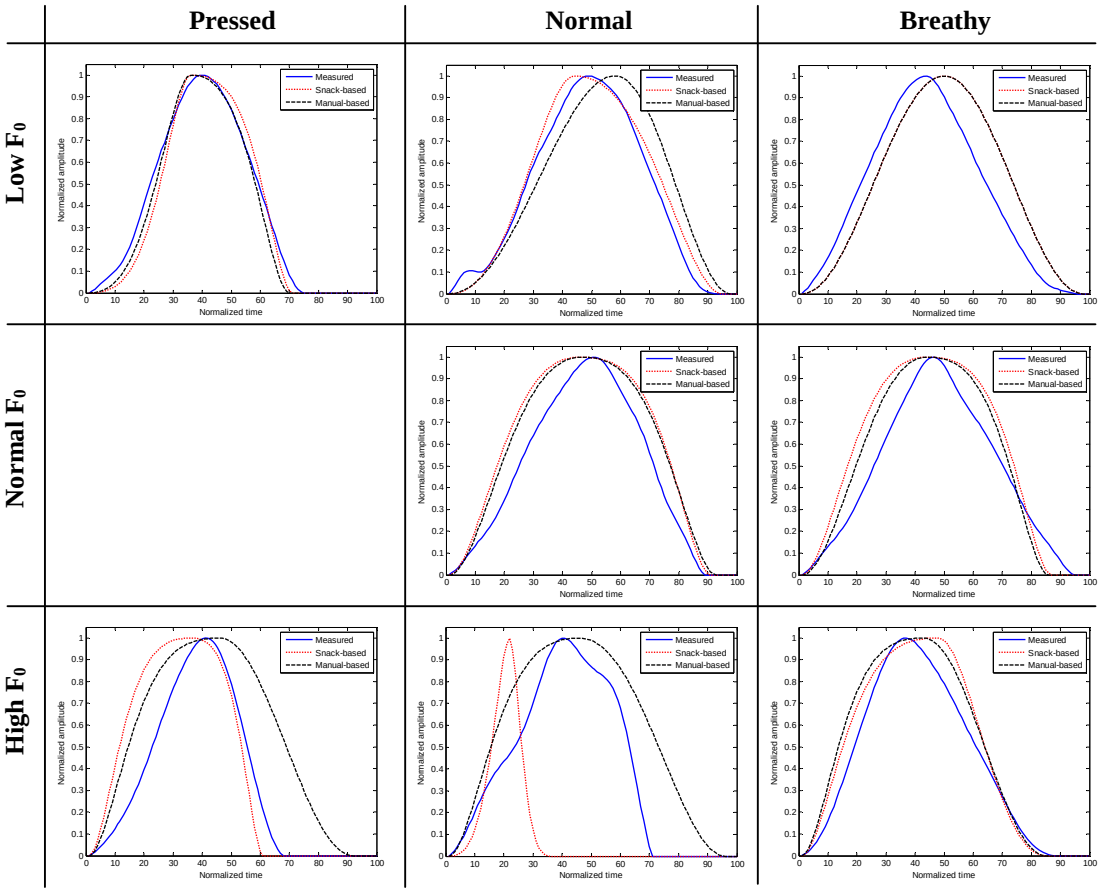



**Figure C.2:** Plot of the measured (solid line) and estimated glottal area waveforms for subject FM2. The estimated waveforms are from the Snack-based (dotted line) and manual-based (dashed line) formant constraints.

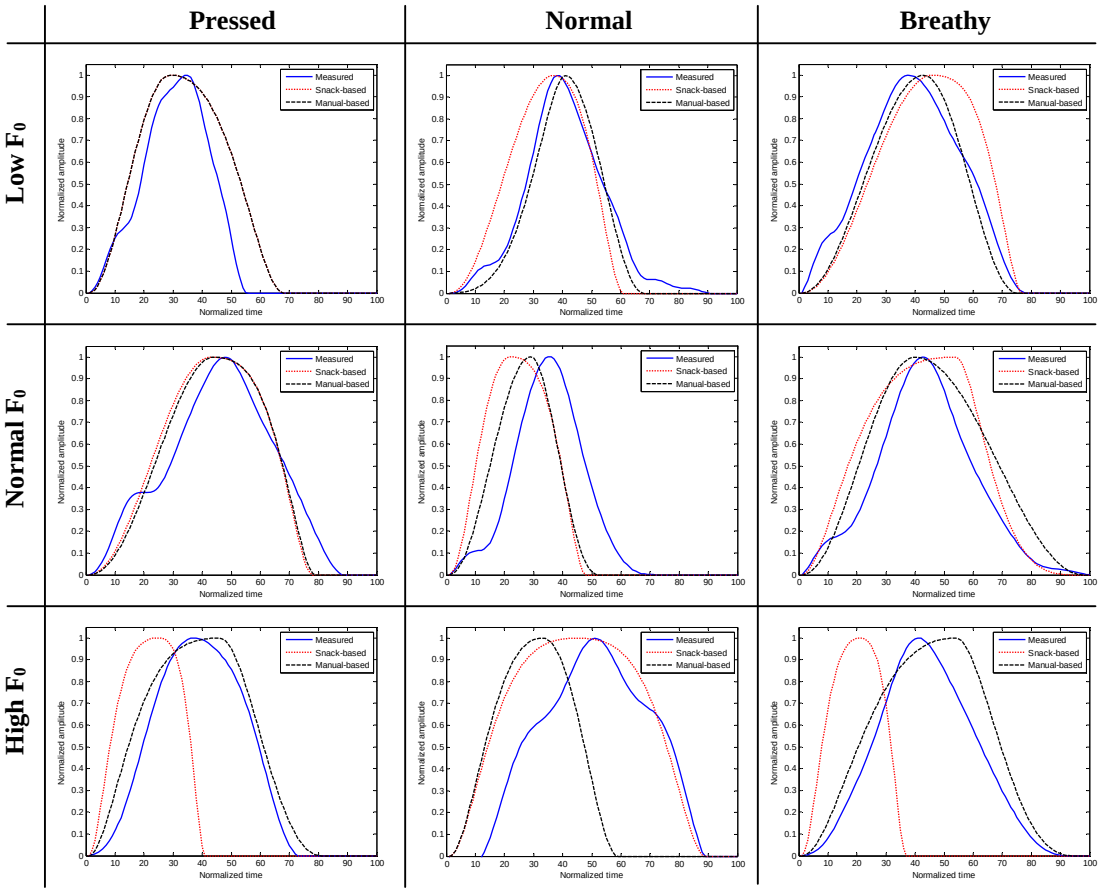



**Figure C.3:** Plot of the measured (solid line) and estimated glottal area waveforms for subject FM3. The estimated waveforms are from the Snack-based (dotted line) and manual-based (dashed line) formant constraints.

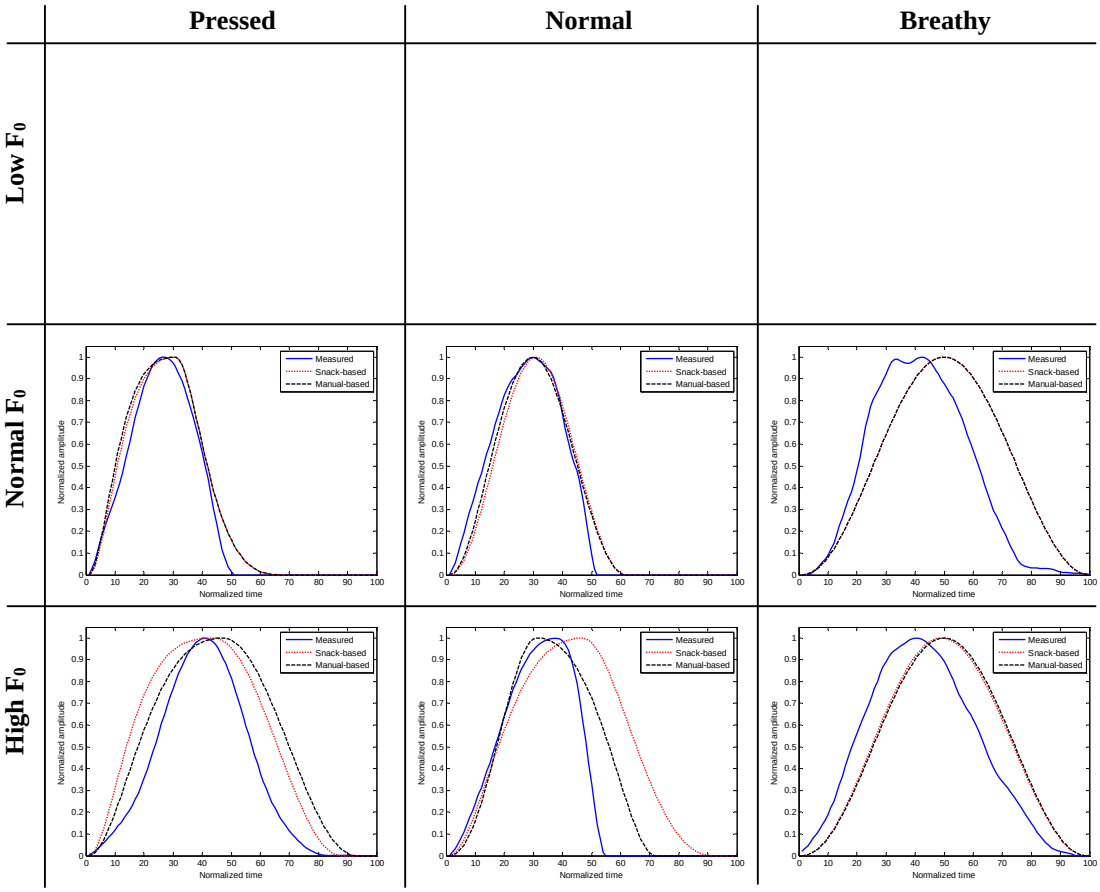



**Figure C.4:** Plot of the measured (solid line) and estimated glottal area waveforms for subject M1. The estimated waveforms are from the Snack-based (dotted line) and manual-based (dashed line) formant constraints.

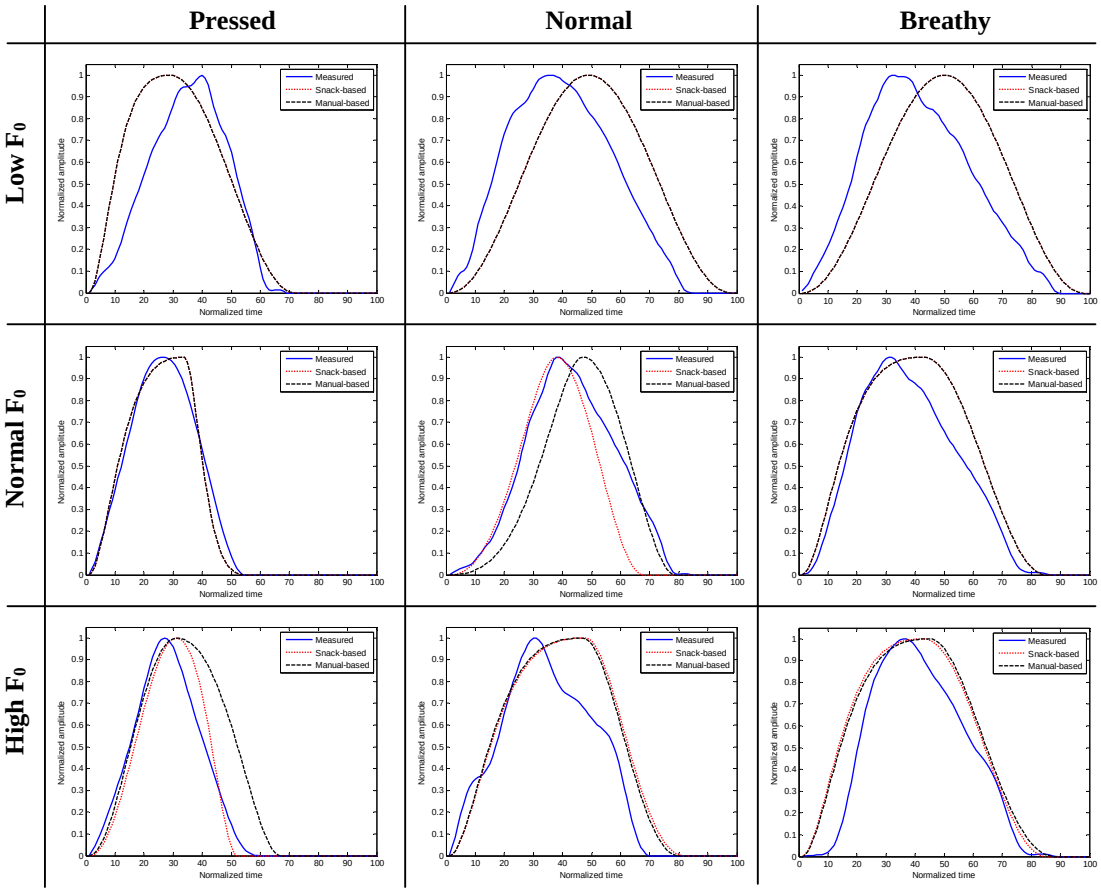



**Figure C.5:** Plot of the measured (solid line) and estimated glottal area waveforms for subject M2. The estimated waveforms are from the Snack-based (dotted line) and manual-based (dashed line) formant constraints.

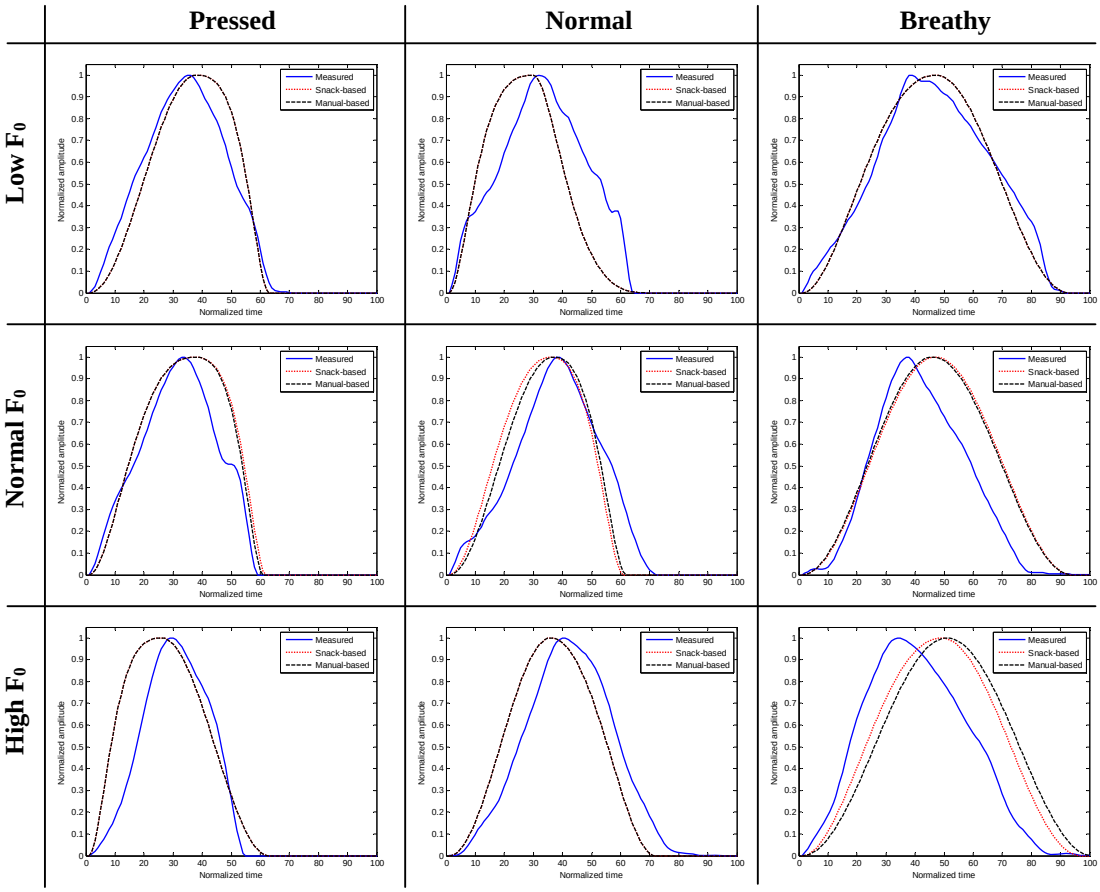



**Figure C.6:** Plot of the measured (solid line) and estimated glottal area waveforms for subject M3. The estimated waveforms are from the Snack-based (dotted line) and manual-based (dashed line) formant constraints.

# References


[AF82] T.V. Ananthapadmanabha and G. Fant. "Calculation of true glottal flow and its components." *Speech Comm.*, **1**:167–184, 1982.

[AG07] A. Arvaniti and G. Garding. "Dialectal variation in the rising accents of American English." *J. Cole and J.H. Hualde (Eds), Papers in Laboratory Phonology 9*, pp. 547–576, 2007.

[ALM06] A. Arvaniti, D.R. Ladd, and I. Mennen. "Phonetic effects of focus and "tonal crowding" in intonation: evidence from Greek polar questions." *Speech Comm.*, **48**:667–696, 2006.

[Ana84] T.V. Ananthapadmanabha. "Acoustic analysis of voice source dynamics." *STL-QPSR*, **25**(2–3):1–24, 1984.

[Ave04] H. Avelino. *Topics in Yalalag Zapotec, with particular reference to its phonetic structures*. PhD thesis, University of California, Los Angeles, 2004.

[BE94] M. Beckman and J. Edwards. "Articulatory evidence for differentiation of stress categories." *P.A. Keating (Ed), Papers in Laboratory Phonology III*, pp. 7–33, 1994.

[Ben80] S. Bennett. "Vowel formant frequency characteristics of preadolescent males and females." *J. Acoust. Soc. Am.*, **67**(S1):S25–S26, April 1980.

[Bic82] C. Bickley. *Acoustic analysis and perception of breathy vowels*. Speech Communication Group Working Papers I, Cambridge, MA: Research Laboratory of Electronics, 1982.

[Bla97] B. Blankenship. *The time course of breathiness and laryngealization in vowels*. PhD thesis, University of California, Los Angeles, 1997.

[BP86] M. Beckman and J.B. Pierrehumbert. "Intonational structure in Japanese and English." *Phonol. Yearbook 3*, pp. 255–309, 1986.

[BP95] P. Busby and G. Plant. "Formant frequency values of vowels produced by preadolescent boys and girls." *J. Acoust. Soc. Am.*, **97**(4):2603–2606, April 1995.

[BW10] P. Boersma and D. Weenink. *Praat: doing phonetics by computer (Version 5.1.25)*, 2010. Software available at `http://www.praat.org`.

[CC95] K.E. Cummings and M.A. Clements. "Glottal models for digital speech processing: a historical survey and new results." *Digital signal processing*, **5(1)**:21–42, 1995.

[CG97] A.N. Chasaide and G. Gobl. "Voice source variations." In *W.J. Hardcastle and J. Laver (Eds), The Handbook of Phonetic Sciences*, pp. 427–461, Blackwell Publishers Inc., Oxford, 1997.

[CH96] S. Cassidy and J. Harrington. "EMU: an enhanced hierarchical speech database management system." In *Proceedings of the 6th Australian International Conference on Speech Science and Technology*, pp. 361–366, 1996.

[Cha93] W.L. Chafe. "Prosodic and functional units of language." *J.A. Edwards and M.D. Lampert (Eds), Talking Data: Transciption and Coding in Discourse Research*, pp. 3–43, 1993.

[CHC05] J.-Y. Choi, M. Hasegawa-Johnson, and J. Cole. "Finding intonational boundaries using acoustic cues related to the voice source." *J. Acoust. Soc. Am.*, **118**(4):2579–2587, October 2005.

[CL01] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[DBP07] M.R. Draper, B. Blagnys, and D.J. Premachandra. "To 'EE' or not to 'EE'." *J. Otolaryngol*, **36**:189–193, 2007.

[DSO96] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf. "Glottalization of vowel-initial syllables as a function of prosodic structure." *J. Phonetics*, **24**:423–444, 1996.

[EM91] A. El-Jaroudi and J. Makhoul. "Discrete all-pole modeling." *IEEE Trans. on Sig. Proc.*, **39**(2):441–423, 1991.

[Eps02] M. Epstein. *Voice quality and prosody in English.* PhD thesis, University of California, Los Angeles, 2002.

[EPY09] C.M. Esposito, J. Ptacek, and S. Yang. "An acoustic and electroglottographic study of White Hmong phonation." In *J. Acoust. Soc. Am.*, p. 2223, San Antonio, TX, 2009.

[Esp03] C.M. Esposito. *Phonation in Santa Ana del Valle Zapotec.* PhD thesis, University of California, Los Angeles, 2003.

[Esp06] C.M. Esposito. *The effects of linguistic experience on the perception of phonation*. PhD thesis, University of California, Los Angeles, 2006.

[Fan70] G. Fant. *Acoustic theory of speech production*. Mouton, The Hague, Paris, 2nd edition, 1970.

[Fan79] G. Fant. "Glottal source and excitation analysis." *STL-QPSR*, **20**(1):85–107, 1979.

[Fan97] G. Fant. "The voice source in connected speech." *Speech Comm.*, **22**(2–3):125–139, 1997.

[Fis67] E. Fischer-Jorgensen. "Phonetic analysis of breathy (murmured) vowels in Gujarati." *Indian Linguist*, **28**:71–139, 1967.

[FJ98] C. Fougeron and S.-A. Jun. "Rate effects on French intonation: prosodic organization and phonetic realization." *J. Phonetics*, **28**:161–185, 1998.

[FL86] H. Fujisaki and M. Ljungqvist. "Proposal and evaluation of models for the glottal source waveform." In *Proceedings of ICASSP*, pp. 1605–1608, Tokyo, Japan, 1986.

[Fla65] J.L. Flanagan. *Speech analysis and perception*. Springer-Verlag, Berlin, 2nd edition, 1965.

[FLL85] G. Fant, J. Liljencrants, and Q. Lin. "A four-parameter model of glottal flow." *STL-QPSR*, pp. 1–14, 1985.

[FMS01] M. Fröhlich, D. Michaelis, and H.W. Strube. "SIM – simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals." *J. Acoust. Soc. Am.*, **110**:479–488, 2001.

[GPN00] E. Grabe, B. Post, F. Nolan, and K. Farrar. "Pitch accent realization in four varieties of British English." *J. Phonetics*, **28**:161–185, 2000.

[Han97] H.M. Hanson. "Glottal characteristics of female speakers: Acoustic correlates." *J. Acoust. Soc. Am.*, **101**:466–481, 1997.

[HC99] H.M. Hanson and E.S. Chuang. "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data." *J. Acoust. Soc. Am.*, **106**:1064–1077, 1999.

[HCE94] J. Hillenbrand, R.A. Cleveland, and R.L. Erickson. "Acoustic correlates of breathy vocal quality." *J. Speech and Hearing Research*, **37**:769–778, 1994.

[HdD01] N. Henrich, C. d'Alessandro, and B. Doval. "Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data." In *Proceedings of EUROSPEECH*, pp. 47–50, Aalborg, Denmark, 2001.

[Hed84] P. Hedelin. "A glottal LPC-vocoder." In *Proceedings of ICASSP*, pp. 161–164, San Diego, CA, 1984.

[HG92] S. Hertegård and J. Gauffin. "Acoustic properties of the Rothenberg mask." *STL-QPSR*, **33**(2–3):9–18, 1992.

[HHP95] E.B. Holmberg, R.E. Hillman, J.S. Perkell, P. Guiod, and S.L. Goldman. "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice." *J. Speech Hear. Res.*, **38**:1212–1223, 1995.

[HM95] J.W. Hawks and J.D. Miller. "A formant bandwidth estimation procedure for vowel synthesis." *J. Acoust. Soc. Am.*, **97**:1343–1344, 1995.

[HM07] M. Howe and R. McGowan. "Sound generated by aerodynamic sources near a deformable body, with application to voiced speech." *J. Fluid Mech.*, **592**:367–392, 2007.

[HP86] J. Hirschberg and J. Pierrehumbert. "The intonational structuring of discourse." In *Proceedings of 24th Annual Meeting on Association for Computational Linguistics*, pp. 136–144, 1986.

[Huf87] M.K. Huffman. "Measures of phonation type in Hmong." *J. Acoust. Soc. Am.*, **81**(2):495–504, February 1987.

[IA04] M. Iseli and A. Alwan. "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation." In *Proceedings of ICASSP*, volume 1, pp. 669–672, Montreal, Canada, May 2004.

[IC04] E. Moore II and M. Clements. "Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information." In *Proceedings of ICASSP*, pp. 101–104, Montreal, Canada, 2004.

[IS68] F. Itakura and S. Saito. "Analysis synthesis telephony based on the maximum likelihood method." In *Proceedings of 6th Int. Congr. Acoust.*, pp. C17–C20, Tokyo, Japan, 1968.

[ISA07] M. Iseli, Y.-L. Shue, and A. Alwan. "Age, sex, and vowel dependencies of acoustic measures related to the voice source." *J. Acoust. Soc. Am.*, **121**(4):2283–2295, 2007.

[ISE06] M. Iseli, Y.-L. Shue, M. Epstein, P. Keating, J. Kreiman, and A. Alwan. "Voice source correlates of prosodic features in American English: A pilot study." In *Proceedings of Interspeech*, pp. 2226–2229, Pittsburgh, PA, September 2006.

[Ise07] M. Iseli. *Dependencies of voice source measures on age, sex, vowel context, and prosodic features*. PhD thesis, University of California, Los Angeles, 2007.

[JI05] P. Jinachitra and J.O. Smith III. "Joint estimation of glottal source and vocal tract for vocal synthesis using Kalman smoothing and EM algorithm." In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 327–330, 2005.

[JM07] M. Jilka and B. Möbius. "The influence of vowel quality features on peak alignment." In *Proceedings of Interspeech*, pp. 2621–2624, Antwerp, Belgium, 2007.

[KCP98] H. Kawahara, A. de Cheveigné, and R.D. Patterson. "An instantaneous-frequency-based pitch extraction method for high quality speech transformation: revised TEMP in the STRAIGHT-suite." In *Proceedings of ICSLP*, Sydney, Australia, 1998.

[KGB07] J. Kreiman, B.R. Gerratt, and N. Anto nanzas Barroso. "Measures of the glottal source spectrum." *J. Speech, Language, and Hearing Research*, **50**:595–610, June 2007.

[KGC05] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner. "Loudness predicts prominence: fundamental frequency lends little." *J. Acoust. Soc. Am.*, **118**(2):1038–1054, August 2005.

[KGI08] J. Kreiman, B.R. Gerrat, M. Iseli, J. Neubauer, Y.-L. Shue, and A. Alwan. "The relationship between open quotient and $H_1^* - H_2^*$." In *Proceedings of 6th International Conf. on Voice Physiology and Biomechanics*, Tampere, Finland, Aug. 2008.

[Kha09] S. Khan. "An acoustic and electroglottographic study of breathy phonation in Gujarati." In *J. Acoust. Soc. Am.*, p. 2222, San Antonio, TX, 2009.

[KK90] D.H. Klatt and L.C. Klatt. "Analysis, synthesis, and perception of voice quality variations among female and male talkers." *J. Acoust. Soc. Am.*, **87**(2):820–857, February 1990.

[Kla76a] D.H. Klatt. "Linguistic uses of segmental duration in English: acoustic and perceptual evidence." *J. Acoust. Soc. Am.*, **59**(5):1208–1221, May 1976.

[Kla76b] D.H. Klatt. "Structure of a phonological rule component for a synthesis-by-rule program." *IEEE Trans. Acoust. Speech Signal Process*, **24**(5):391–398, October 1976.

[Kro93] G. de Krom. "A cepstrum-based technique for determining a harmonic-to-noise ratio in speech signals." *J. Speech and Hearing Research*, **36**:254–266, 1993.

[Lad08] D.R. Ladd. *Intonational Phonology*. Cambridge University Press, Cambridge, 1996/2008.

[Lev] S.V. Levi. "Intonation in Turkish: the realization of noun compounds and genitive possessive NPs." *submitted for publication*.

[LPN99] S. Lee, A. Potamianos, and S. Narayanan. "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters." *J. Acoust. Soc. Am.*, **105**(3):1455–1468, March 1999.

[Mak75a] J. Makhoul. "Linear prediction: A tutorial review." *Proceedings of the IEEE*, **63**(4):561–580, 1975.

[Mak75b] J. Makhoul. "Spectral linear prediction: Properties and applications." *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-23**(3):283–296, 1975.

[MGB09] D. Mücke, M. Grice, J. Becker, and A. Hermes. "Sources of variation in tonal alignment: evidence from acoustic and kinematic data." *J. Phonetics*, **37**(3):321–338, 2009.

[MLU96] J.D. Miller, S. Lee, R.M. Uchanski, A.F. Heidbreder, B.B. Richman, and J. Tadlock. "Creation of two children's speech databases." In *Proceedings of ICASSP*, volume 2, pp. 849–852, Atlanta, Georgia, May 1996.

[Od05] C. Odé. "Neutralization or truncation? The perception of two Russian pitch accents on utterance-final syllables." *Speech Comm.*, **47**(1–2):71–79, 2005.

[OE73] J.J Ohala and W.G. Ewan. "Speed of pitch change (A)." *J. Acoust. Soc. Am.*, **53**(1):345, 1973.

[Oko06] A. Okobi. *Acoustic correlates of word stress in American English*. PhD thesis, Massachusetts Institute of Technology, 2006.

[PB52] G.E. Peterson and H.L. Barney. "Control methods used in a study of the vowels." *J. Acoust. Soc. Am.*, **24**(2):175–184, March 1952.

[PB09] J. Pérez and A. Bonafonte. "Towards robust glottal source modeling." In *Proceedings of Interspeech*, pp. 68–71, Brighton, UK, 2009.

[Pie80] J.B. Pierrehumbert. *The phonology and phonetics of English intonation*. PhD thesis, Massachusetts Institute of Technology, 1980.

[POA01] T.L. Perry, R.N. Ohde, and D.H. Ashmead. "The acoustic bases for gender identification from children's voices." *J. Acoust. Soc. Am.*, **109**(6):2988–2998, June 2001.

[PT91] J.B. Pierrehumbert and D. Talkin. "Lenition of /h/ and glottal stop." *Papers in Lab. Phon. II*, pp. 90–117, 1991.

[PY08] A. del Pozo and S. Young. "The linear transformation of LF glottal waveforms for voice conversion." In *Proceedings of Interspeech*, pp. 1457–1460, Brisbane, Australia, 2008.

[RH06] A. Rosenberg and J. Hirschberg. "On the correlation of energy and pitch accent in read English speech." In *Proceedings of Interspeech*, pp. 301–304, Pittsburgh, PA, 2006.

[Ros71] A. Rosenberg. "Effects of the glottal pulse shape on the quality of natural vowels." *J. Acoust. Soc. Am.*, **49**(2):583–590, 1971.

[Rot73] M. Rothenberg. "A new inverse-filtering technique for deriving the glottal airflow during voicing." *J. Acoust. Soc. Am.*, **53**(6):1632–1645, June 1973.

[SBP92] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, and J. Pierrehumbert. "ToBI: a standard for labeling English prosody." In *Proceedings of ICSLP*, volume 2, pp. 867–870, Alberta, Canada, October 1992.

[Sj04] Kåre Sjölander. "Snack Sound Toolkit." KTH Stockholm, Sweden, 2004. `http://www.speech.kth.se/snack/` (last viewed Aug. 2009).

[SL90] M. Södersten and P.-A. Lindestad. "Glottal closure and perceived breathiness during phonation in normally speaking subjects." *J. Speech and Hearing Research*, **33**:601–611, 1990.

[Sli07] J. Slifka. "Some physiological correlates to regular and irregular phonation at the end of an utterance." *J. Voice*, **20**:171–186, 2007.

[SP90] K.E.A. Silverman and J.B. Pierrehumbert. "The timing of prenuclear high accents in English." In *Papers in Lab. Phon. I*, pp. 72–106, Cambridge University Press, Cambridge, UK, 1990.

[ST96] S. Shattuck-Hufnagel and A. Turk. "A prosody tutorial for investigators of auditory sentence processing." *J. of Psycholinguistic Research*, **25**(2):193–247, 1996.

[Ste00] K.N. Stevens. *Acoustic phonetics*. The MIT Press, 2000.

[Sun79] J. Sundberg. "Maximum speed of pitch changes in singers and untrained subjects." *J. Phonetics*, **7**:71–79, 1979.

[SV96a] A.M.C. Sluijter and V.J. Van Heuven. "Acoustic correlates of linguistic stress and accent in Dutch and American English." In *Proceedings of ICSLP*, pp. 630–633, Philadelphia, PA, 1996.

[SV96b] A.M.C. Sluijter and V.J. Van Heuven. "Spectral balance as an acoustic correlate of linguistic stress." *J. Acoust. Soc. Am.*, **100**(4):2471–2485, 1996.

[Tit89] I.R. Titze. "Physiologic and acoustic difference between male and female voices." *J. Acoust. Soc. Am.*, **85**(4):1699–1707, 1989.

[TS07] A.E. Turk and S. Shattuck-Hufnagel. "Phrase-final lengthening in American English." *J. Phonetics*, **35**(4):445–472, 2007.

[TW99] A.E. Turk and L. White. "Structural effects on pitch accentual lengthening in English." *J. Phonetics*, **27**(2):171–206, April 1999.

[WB71] B. Weinberg and S. Bennett. "Speaker sex recognition of 5- and 6-year-old children's voices." *J. Acoust. Soc. Am.*, **50**(4B):1210–1213, October 1971.

[WC91] K. Wu and D.G. Childers. "Gender recognition from speech. Part I: Coarse analysis." *J. Acoust. Soc. Am.*, **90**(4):1828–1840, October 1991.

[WJ03] R. Wayland and A. Jongman. "Acoustic correlates of breathy and clear vowels: the case of Khmer." *J. Phonetics*, **31**:181–201, 2003.

[XS02] Y. Xu and X. Sun. "Maximum speed of pitch change and how it may relate to speech." *J. Acoust. Soc. Am.*, **111**(3):1399–1413, 2002.

[Xu97] Y. Xu. "Contextual tonal variation in Mandarin Chinese." *J. Phonetics*, **25**:619–83, 1997.