A Sociophonetic Investigation of Ethnolinguistic Differences in Voice Quality Among Young, South African English Speakers

Bruce Rory Wileman

Thesis Presented for the Degree of DOCTOR OF PHILOSOPHY in the School of African and Gender Studies, Anthropology and Linguistics

UNIVERSITY OF CAPE TOWN

March 2018

The financial assistance of the National Research Foundation (NRF grant number 64805) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

ABSTRACT

Thesis Title: A sociophonetic investigation of ethnolinguistic differences in voice quality among young, South African English speakers.

Author: Bruce Rory Wileman

Prior research has suggested that there may be differences in voice quality between black and white speakers of South African English who had attended well-resourced middle-class schools. The principal objective of the study is to address the question of whether there is any acoustic evidence of such differences. The study then proceeds to describe such acoustic evidence for differences in voice quality.

The author interviewed 36 female South African English speakers (18 white and 18 black) between the ages of 18 and 22. The research subjects had all attended well-resourced middle-class schools. In order to control for the possibility of substrate influences on voice quality, all black participants were of an isiXhosa language background. High quality sound recordings were conducted, consisting of both a set of read sentences as well as semi-structured interviews, the latter of which formed the core dataset for the subsequent acoustic analysis. The acoustic data were analyzed using VoiceSauce, a program specifically designed for the acoustic analysis of voice quality. Measurements were based on automatically segmented speech samples using FAVE and PRAAT. The VoiceSauce measurement data were statistically analyzed by means of a linear mixed effects regression analysis and Wilcoxon rank sum tests using the statistical package R to evaluate the significance of ethnicity as a variable.

The effect of ethnicity was found to be significant for several measures of spectral tilt (including for example, 2K*-5K, H4*-2K*, H1*-H2* and H1*-A1*) and cepstral peak prominence with a nearly significant effect for the subharmonics-to-harmonics ratio. Black speakers exhibited consistently higher values for most harmonic differential measures (for example, H1*-A1*) overall, while white speakers exhibited higher values for fundamental frequency, harmonics-to-noise ratio and cepstral peak prominence. The author concludes that the acoustic evidence is most consistent with the hypothesis that the white speakers overall typically use a voice quality

characterized by greater vocal fold constriction, thickness and stiffness in comparison to the black speakers, hypothesized to use a voice quality characterized by more breathiness. By providing a description of voice quality variation, the research contributes towards a more complete account of sociolinguistic variation in South African English.

ACKNOWLEDGEMENTS

I would like to acknowledge the following individuals and organizations for their contributions towards this research project. Firstly, I would like to thank Rajend Mesthrie, without whose funding, support, advice, generosity, patience and willingness to take on the supervision of this thesis, this project could never have materialized. I would like to acknowledge the funding contribution of the National Research Foundation, via Professor Mesthrie's research chair in Migration, Language and Social Change. I also appreciate the funding assistance I received from UCT via the Lestrade scholarship. I would also like to thank my advisory supervisor, Daan Wissing, for providing much needed advice based on his expertise as well as for the assurance and encouragement. Many thanks also to Wikus Pienaar and Daniel van Niekerk in particular, for their advice on the speech recordings. I would also like to express my gratitude to Kirsten Morreira and Roger Lass for their helpful feedback on some of my earlier work. Thank you also to Tracey Toefy for her assistance particularly in terms of being prepared to avail herself for providing statistical advice. Many thanks to Yolandi Klein for her assistance in the auditory analysis for this project. Thank you to Mastin Prinsloo as well as Alida Chevalier, for their help in organizing research subjects. I would also like to thank Alida Chevalier for her help in proof-reading some of my work. A great many thanks to Alan Johannes for his patience and assistance at the recording studio, as well as Molly Maunganidze. Thank you also to Sven Oxtoby, for his help in testing the sound equipment. Thank you to Julia Laurie and Alicia Kamaldien for their assistance with transcription. A special thank you to research assistant Alicia Kamaldien for going beyond the call of duty in terms of transcription, segmentation and re-alignment work. Thank you to both Faiza Steffenson and Alida Chevalier for general administrative duties and keeping things running smoothly in the Linguistics Section at UCT. Thank you to Deon du Plessis and Kara Schultz for their insights. Thank you to both Yen Liang-Shue and Pat Keating, for their help, patience and advice regarding VoiceSauce. Thank you also to Pat Keating for feedback on my chapters. Thank you also to Erez Levon, Ian Bekker, as well as Andries Coetzee for invaluable and often very detailed feedback. Finally, this project would not have been possible were it not for my research subjects. A great many thanks to you all. Thank you to my family for their support. My apologies if I have left anyone out.

PLAGIARISM DECLARATION

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

2. I have used the Journal of Sociolinguistics convention for citation and referencing. Each significant contribution to, and quotation in, this thesis from the works of other people has been attributed, and has been cited and referenced.

3. This thesis is my own work.

4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

5. I acknowledge that copying someone else's work, or part of it, is wrong, and declare that this is my own work.

Signature: Signed by candidate Date: 01/08/2017

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGEMENTS iv
PLAGIARISM DECLARATION v
TABLE OF CONTENTS vi
LIST OF FIGURES xiii
LIST OF TABLES xvii
LIST OF ABBREVIATIONS xvii
Chapter I: INTRODUCTION 1
1.1. BACKGROUND 1
1.2. VOICE QUALITY THEORY: A BRIEF INTRODUCTION 4
1.3. MORE RECENT DEVELOPMENTS IN THE STUDY OF VOICE QUALITY 10
1.4. VOICE QUALITY DEFINED AND OPERATIONALIZED 13
1.5. FORMAL PROBLEM STATEMENTS 16
1.6. FORMAL STATEMENT OF RESEARCH OBJECTIVES 18
1.7. THESIS STRUCTURE 18
1.8. CONCLUSION 19
CHAPTER II: BACKGROUND 20
2.1. INTRODUCTION 20
2.2. VOICE QUALITY AS AN INDEX 20
2.3. THE INVESTIGATION OF THE SOCIOLINGUISTIC FUNCTION OF VOICE QUALITY VARIATION 25
2.3.1. Voice Quality as an Index of Regional Provenance 27
2.3.2. Voice Quality as an Index of Social Class Stratification within Regional Dialects 28
2.3.3. Voice Quality and Gender 32
2.4. STUDIES OF ETHNIC DIFFERENCES IN VOICE QUALITY 34
2.4.1. Anatomical Differences vs Learned Behaviour 35
2.4.2. Physiological Studies 36
2.4.3. Perception Studies 37
2.5. ACOUSTIC STUDIES OF ETHNIC DIFFERENCES IN VOICE QUALITY 41

2.6. VOICE QUALITY VARIATION IN SOUTH AFRICAN ENGLISH (SAE): A REVIEW OF THE RELEVANT LITERATURE 45
2.7. VOICE QUALITY IN ISIXHOSA 49
2.8. REVIEW OF THE ACOUSTIC MEASURES USED IN THIS STUDY 52
Harmonic Differential Measures 53
2.8.1. H1−H2 53
2.8.2. H2−H4 59
2.8.3. H1−A1 63
2.8.4. H1−A2 69
2.8.5. H1−A3 71
2.8.6. H4−H2K 73
2.8.7. 2K−5K 74
2.8.8. Measures of Signal Aperiodicity in VS 75
2.8.8.1. HNR (Harmonics-to-Noise Ratio) 76
2.8.8.2. Cepstral Peak Prominence (CPP) 79
2.8.8.3. Subharmonic-to-Harmonic Ratio (SHR) 83
2.8.9. Energy (Root Mean Square Energy) 84
2.9. THE PSYCHOACOUSTIC MODEL 85
2.10. CONCLUSION 89
CHAPTER III: METHODOLOGY 90
3.1. INTRODUCTION 90
3.2. SAMPLING 90
3.2.1. Sample Collection 91
3.2.2. Sample Characteristics 92
3.3. RECORDING PROCEDURE 92
3.4. ACOUSTIC DATA ANALYSIS USING VOICESAUCE (VS) 94
3.4.1. Initial Data Preparation 94
3.4.2. Annotation in PRAAT using PRAAT Text Grids 95
3.4.3. Automatic Segmentation using FAVE (Forced Alignment and Vowel Token Segment Extraction) 96
3.4.4. Segmentation Checking and Extraction using PRAAT 96
3.5. VS PARAMETER DESCRIPTIONS 97
3.5.1. Harmonic Differential Measures 97
3.5.1.1. H1−H2 98

3.5.1.2. H2−H4 98
3.5.1.3. H1−A1 99
3.5.1.4. H1−A2 100
3.5.1.5. H1−A3 100
3.5.1.6. H4−H2K 100
3.5.1.7. H2K−5K 101
3.5.2. Measures of Signal Aperiodicity and Noise in VS 101
3.5.2.1. HNR (Harmonics-to-Noise Ratio) 101
3.5.2.2. Cepstral Peak Prominence (CPP) 101
3.5.2.3. Subharmonic-to-Harmonic Ratio (SHR) 102
3.5.3. Energy (Root Mean Square Energy) 102
3.6. GENERAL VS SETTINGS 103
3.7. ACOUSTIC MEASUREMENT OF FORMANT FREQUENCIES USING THE PRAAT ALGORITHM IN VS 104
3.8. AUDITORY ANALYSIS 105
3.9. COMPARISON AND STATISTICAL ANALYSIS USING LINEAR MIXED EFFECTS REGRESSION AND WILCOXON RANK SUM TESTS IN R 114
3.10. CONCLUSION 121
CHAPTER IV: RESULTS FOR THE AUDITORY ANALYSIS AND THE HARMONIC DIFFERENTIAL MEASURES OF THE ACOUSTIC ANALYSIS 122
4.1. INTRODUCTION 122
4.2. AUDITORY ANALYSIS OF THE SENTENCE DATA 123
4.3. ACOUSTIC DATA ANALYSIS 130
4.3.1. Measures of spectral balance 130
4.3.1.1. 2K*−5K (the amplitude of the harmonic nearest to 5000 Hz subtracted from the harmonic nearest 2000 Hz, the latter of which is corrected for the effect of formants and their bandwidths) 130
4.3.1.1.1. 2K*−5K Sentence Data and the Auditorily Identified Phonation Types 130
4.3.1.1.2. 2K*−5K Statistical Analysis 131
4.3.1.1.2.1. Interview Data 131
4.3.1.1.2.2. Sentence Data Linear Mixed Effects Analysis Results 134
4.3.1.1.2.3. Discussion 134
4.3.1.1.3. Summary of the Findings for the 2K*−5K Measure 139
4.3.1.1.4. Summary of Results for 2K−5K (the uncorrected equivalent measure of 2K*−5K) 139

4.3.1.2. H4*−2K* (the fourth harmonic minus the amplitude of the harmonic nearest 2000 Hz, both corrected for the influence of formants and their bandwidths) 140
4.3.1.2.1. H4*−2K* Sentence Data and the Auditorily Identified Phonation Types 140
4.3.1.2.2. H4*−2K* Statistical Analysis 141
4.3.1.2.2.1. Interview Data 141
4.3.1.2.2.2. Sentence Data Linear Mixed Effects Analysis Results 143
4.3.1.2.2.3. Discussion 143
4.3.1.2.3. Summary of the Findings for H4*−2K* 148
4.3.1.2.4. Summary of results for H4−2K (the uncorrected equivalent of H4*−2K*) 148
4.3.1.3. H1*−H2* (The first harmonic minus the second harmonic, both corrected for the influence of formants and their amplitudes) 149
4.3.1.3.1. H1*−H2* Sentence Data and the Auditorily Identified Phonation Types 149
4.3.1.3.2. Statistical Analysis for H1*−H2* 150
4.3.1.3.2.1. Interview Data 150
4.3.1.3.2.2. Sentence Data Linear Mixed Effects Analysis Results 153
4.3.1.3.2.3. Discussion 153
4.3.1.3.3. Summary of Findings for H1*−H2* 158
4.3.1.4. H1−H2 (the uncorrected equivalent measure of H1*−H2*) 159
4.3.1.4.1. H1−H2 Sentence Data and the Auditorily Identified Phonation Types 159
4.3.1.4.2. Statistical Analysis for H1−H2 160
4.3.1.4.2.1. Interview Data 160
4.3.1.4.2.2. Sentence Data Linear Mixed Effects Analysis Results 163
4.3.1.4.3. Summary of Findings for H1−H2 163
4.3.1.5. H2*−H4* (the second harmonic minus the fourth harmonic, both corrected for the influence of formants and their amplitudes) 163
4.3.1.5.1. H2*−H4* Sentence Data and the Auditorily Identified Phonation Types 163
4.3.1.5.2. Statistical Analysis for H2*−H4* 164
4.3.1.5.2.1. Interview Data 166
4.3.1.5.2.2. Sentence Data Linear Mixed Effects Analysis Results 166
4.3.1.5.3. Summary of Findings for H2*−H4* 166
4.3.1.5.4. Summary of Findings for H2−H4 (the uncorrected equivalent of H2*−H4*) 167
4.3.2. Other Spectral Amplitude Measures 168
4.3.2.1. H1*−A1* (the amplitude of the first harmonic minus the amplitude of the harmonic nearest F1, both corrected for the influence of formants and their bandwidths) 168

4.3.2.1.1. H1*−A1* Sentence Data and the Auditorily Identified Phonation Types 168
4.3.2.1.2. Statistical Analysis for H1*−A1* 169
4.3.2.1.2.1. Interview Data 170
4.3.2.1.2.2. Sentence Data Linear Mixed Effects Analysis Results 171
4.3.2.1.2.3. Discussion 171
4.3.2.1.3. Summary of Findings for H1*−A1* 175
4.3.2.1.4. Summary of Findings for H1−A1 (the uncorrected equivalent measure of H1*−A1*) 175
4.3.2.2. H1*−A2* (the amplitude of the first harmonic minus the amplitude of the harmonic nearest F2, both corrected for the influence of formants and their bandwidths) 176
4.3.2.2.1. H1*−A2* Sentence Data and the Auditorily Identified Phonation Types 176
4.3.2.2.2. Statistical Analysis for H1*−A2* 177
4.3.2.2.2.1. Interview Data 178
4.3.2.2.2.2. Sentence Data Linear Mixed Effects Analysis Results 179
4.3.2.2.3. Summary of Findings for H1*−A2* 179
4.3.2.2.4. Summary of Findings for H1−A2 (the uncorrected equivalent of H1*−A2*) 180
4.3.2.3. H1*−A3* (the amplitude of the first harmonic minus the amplitude of the harmonic nearest F3, both corrected for the effects of formants and formant amplitudes) 181
4.3.2.3.1. H1*−A3* Sentence Data and the Auditorily Identified Phonation Types 181
4.3.2.3.2. Statistical Analysis for H1*−A3* 182
4.3.2.3.2.1. Interview Data 184
4.3.2.3.2.2. Sentence Data Linear Mixed Effects Analysis Results 184
4.3.2.3.3. Summary of Findings for H1*−A3* 184
4.3.2.3.4. Summary of Findings for H1−A3 (the uncorrected equivalent of H1*−A3*) 185
4.4. CONCLUSION 186
CHAPTER V: RESULTS OF THE ACOUSTIC ANALYSIS FOR THE NOISE MEASURES AND AN OVERVIEW OF THE RESEARCH FINDINGS 187
5.1. INTRODUCTION 187
5.2. SHR (SUBHARMONICS-TO-HARMONICS RATIO) 187
5.2.1. SHR Sentence Data and the Auditorily Identified Phonation Types 187
5.2.2. Statistical Analysis for SHR 188
5.2.2.1. Interview Data 190
5.2.2.2. Sentence Data Linear Mixed Effects Analysis Results 191
5.2.3. Summary of Findings for SHR 191

5.3. CPP (CEPSTRAL PEAK PROMINENCE) 191
5.3.1. CPP Sentence Data and the Auditorily Identified Phonation Types 191
5.3.2. Statistical Analysis for the CPP Data 192
5.3.2.1. Interview Data 194
5.3.2.2. Sentence Data Linear Mixed Effects Analysis Results 194
5.3.2.3. Discussion 194
5.3.3. Summary of Findings for CPP 204
5.4. HARMONICS-TO-NOISE-RATIO 205
5.4.1. HNR05 (the harmonics-to-noise ratio between 0 Hz and 500 Hz) 205
5.4.1.1. HNR05 Sentence Data and the Auditorily Identified Phonation Types 205
5.4.1.1. Statistical Analysis for HNR05 206
5.4.1.1.1. Interview Data 208
5.4.1.1.2. Sentence Data Linear Mixed Effects Analysis Results 208
5.4.1.2. Summary of the Findings for HNR05 208
5.4.2. HNR15 (the harmonics-to-noise ratio between 0 Hz and 1500 Hz) 209
5.4.2.1. HNR15 Sentence Data and the Auditorily Identified Phonation Types 209
5.4.2.2. Statistical Analysis for HNR15 210
5.4.2.2.1. Interview Data 212
5.4.2.2.2. Sentence Data Linear Mixed Effects Analysis Results 213
5.4.2.3. Summary of the Findings for HNR15 213
5.4.3. HNR25 (the harmonics-to-noise ratio between 0 Hz and 2500 Hz) 213
5.4.3.1. HNR25 Sentence Data and the Auditorily Identified Phonation Types 213
5.4.3.2. Statistical Analysis for HNR25 214
5.4.3.2.1. Interview Data 216
5.4.3.2.2. Sentence Data Linear Mixed Effects Analysis Results 216
5.4.3.3. Summary of Findings for HNR25 216
5.4.4. HNR35 (the harmonics-to-noise ratio between 0 Hz and 3500 Hz) 217
5.4.4.1. HNR35 Sentence Data and the Auditorily Identified Phonation Types 217
5.4.4.2. Statistical Analysis for HNR35 218
5.4.4.2.1. Interview Data 219
5.4.4.2.2. Sentence Data Linear Mixed Effects Analysis Results 220
5.4.4.3. Summary of Findings for HNR35 220
5.5. OVERVIEW OF RESEARCH FINDINGS 220

CHAPTER VI: DISCUSSION, EXPLANATORY HYPOTHESES AND CONCLUSION 224
6.1. INTRODUCTION 224
6.2. DISCUSSION OF HYPOTHESES BASED ON THE RESEARCH FINDINGS 224
6.2.2. Non-constricted Creak 224
6.2.3. The Breathy/lax/slack Versus Pressed/tense/stiff/constricted Voice Hypothesis 231
6.2.4. Hypothesized Differences in the Abruptness of Vocal Fold Closure 234
6.2.5. Hypothesized Nonsimultaneous Glottal Closure 235
6.2.7. Hypothesized Differences Relating to Open Quotient and Glottal Stricture 238
6.2.8. Other Acoustic Evidence Supporting the Breathy/slack/lax Versus Pressed/stiff/tense/constricted Hypothesis 239
6.2.9. Interpretations Based on More Recent Models 242
6.2.10. Indirect Evidence for the Breathy/slack/lax Versus Tense/pressed/constricted/stiff Voice Hypothesis 246
6.2.11. Overall Summary of the Breathy/lax/slack voice Versus Tense/stiff/constricted/pressed Voice Conclusions 249
6.3. POTENTIAL SOCIOLINGUISTIC SIGNIFICANCE AND THE ROLE OF VOICE QUALITY VARIATION AS AN INDEX OF ETHNIC IDENTITY 252
6.3.1. Sociolinguistic Significance 252
6.3.2. Voice Quality and Ethnicity 254
6.4. RESEARCH IMPLICATIONS 261
6.5. LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH 262
6.5. CONCLUSION 267
APPENDIX A: List of Sentences 268
APPENDIX B: Routinely Elicited Interview Questions 273
APPENDIX C: Descriptive Statistics for the Interview Data 275
APPENDIX D: Dialect History 277
APPENDIX E: Scatterplots 278
References: 296

LIST OF FIGURES

Figure 2.1: A comparison of the volume velocity waveforms for modal and breathy voice taken from Klatt and Klatt (1990:822). 54
Figure 2.2: Comparison of spectra for a modal vowel (2C) and for a breathy vowel (3C) taken from Klatt and Klatt (1990:822). 65
Figure 4.1: The graphic representation of the relative proportions of auditorily identified phonation types occurring in the read sentence data divided according to ethnicity (white colouring symbolizes the number of measurement points for each phonation type for white speakers and dark grey colouring the number of data measurement points for each of the phonation types for black speakers). 125
Figure 4.8: Scatterplot of the 2K*−5K data for white speakers plotted against RMS energy (Pearson's *r*= −0.069). 136
Figure 4.9: Scatterplot of the 2K*−5K data for the whole sample plotted against duration in milliseconds (Pearson's *r*= 0.093). 137
Figure 4.10: Scatterplot of the 2K*−5K data for black speakers plotted against duration in milliseconds (Pearson's *r*= 0.127). 138
Figure 4.11: Scatterplot of the 2K*−5K data for white speakers plotted against duration in milliseconds (Pearson's *r*= 0.06). 138
Figure E4: Scatterplot of 2K*−5K data for the whole sample plotted against pF2 measured in Hertz. 279
Figure E5: Scatterplot of 2K*−5K data for the black speakers plotted against pF2 measured in Hertz. 280
Figure E6: Scatterplot of 2K*−5K data for white speakers plotted against pF2 measured in Hertz. 280
Figure 4.12: Boxplots displaying the values for the H4*−2K* sentence data for each of the auditorily identified phonation types for all data measurement points. 141
Figure 4.13: Boxplots representing the values for black and white speakers for the H4*−2K* interview data. 142
Figure 4.14: Boxplots representing the values for H4*−2K* sentence data for all data measurement points according to ethnicity. 142
Figure 4.2: The graphic representation of the relative proportions of auditorily identified nonmodal phonation types occurring in the read sentence data according to ethnicity. B=breathy, C= prototypical creak, CW=whispery creak, F=vocal fry, FW=whispery fry, H=harsh/aperiodic voice, W= whispery. 126
Figure 4.15: Scatterplot of the H4*−2K* data for the sample as whole plotted against pF1 in Hertz (Pearson's *r*= 0.161). 144
Figure 4.16: Scatterplot of the H4*−2K* data for black speakers plotted against pF1 in Hertz (Pearson's *r*= 0.152). 145
Figure 4.17: Scatterplot of the H4*−2K* data for white speakers plotted against pF1 in Hertz (Pearson's *r*= 0.171). 145
Figure 4.3: Boxplots of the distributions for the 2K*−5K data for each of the auditorily identified phonation types for all data measurement points. 131
Figure 4.18: Scatterplot of H4*−2K* values for the whole sample plotted against duration in milliseconds (Pearson's *r*= −0.055). 146
Figure 4.19: Scatterplot of H4*−2K* data for black speakers plotted against duration in milliseconds (Pearson's *r*= −0.067). 147
Figure 4.20: Scatterplot of H4*−2K* data for white speakers plotted against duration in milliseconds (Pearson's *r*= −0.045). 147
Figure 4.21: Boxplots displaying the values for the H1*−H2* sentence data for each of the auditorily identified phonation types for all data measurement points. 150
Figure 4.22: Boxplots representing the values of black and white speakers for the H1*−H2* interview data. 151

Figure 4.23: Boxplots representing the values for H1*−H2* sentence data for all data measurement points according to ethnicity. 152
Figure 4.24: Boxplots representing the values of black and white speakers for the pF0 interview data. 155
Figure 4.25: Scatterplot of H1*−H2* data for the whole sample plotted against pF0 in Hertz (Pearson's *r*= 0.750). 156
Figure 4.26: Scatterplot of H1*−H2* data for the white speakers plotted against pF0 in Hertz (Pearson's *r*= 0.742). 156
Figure 4.27: Scatterplot of H1*−H2* data for the black speakers plotted against pF0 in Hertz (Pearson's *r*= 0.790). 157
Figure E7: Scatterplot for H1*−H2* data for the interview data for the whole sample plotted against pF1 in Hertz. 281
Figure 4.4: Boxplots representing the values of black and white speakers for the 2K*−5K interview data. 132
Figure E8: Scatterplot for H1*−H2* data for the white speakers plotted against pF1 in Hertz. 281
Figure E9: Scatterplot for H1*−H2* data for the black speakers plotted against pF1 in Hertz. 282
Figure E10: Scatterplot of H1*−H2* data for the whole sample plotted against pF2 in Hertz. 282
Figure E11: Scatterplot of H1*−H2* data for the black speakers plotted against pF2 in Hertz. 283
Figure E12: Scatterplot of H1*−H2* data for the white speakers plotted against pF2 in Hertz. 283
Figure E13: Scatterplot of H1*−H2* data for the whole sample plotted against duration in milliseconds. 284
Figure E14: Scatterplot of H1*−H2* data for the black speakers plotted against duration in milliseconds. 284
Figure E15: Scatterplot of H1*−H2* data for the white speakers plotted against duration in milliseconds. 285
Figure E16: Scatterplot of H1*−H2* data for the whole sample plotted against RMS energy. 285
Figure E17: Scatterplot of H1*−H2* data for the black speakers plotted against RMS energy. 286
Figure E18: Scatterplot of H1*−H2* data for the white speakers plotted against RMS energy. 286
Figure 4.5: Values for the 2K*−5K read sentence data according to ethnicity. 133
Figure 4.29: Boxplots representing the values for black and white speakers for the H1−H2 interview data. 161
Figure 4.46: Boxplots displaying the values for the H1*−A3* sentence data for each of the auditorily identified phonation types for all data measurement points. 182
Figure 4.47: Boxplots representing the values for black and white speakers for the H1*−A3* interview data. 183
Figure 4.30: Boxplots representing the values for H1−H2 sentence data for all data measurement points according to ethnicity. 162
Figure 4.31: Boxplots displaying the values for the H2*−H4* sentence data for each of the auditorily identified phonation types for all data measurement points. 164
Figure 4.32: Boxplot comparison for the values of white and black speakers for the H2*−H4* interview data. 165
Figure 4.28: Boxplots displaying the values for the H1−H2 sentence data for each of the auditorily identified phonation types for all data measurement points. 160
Figure 4.33: Boxplots representing the values for H2*−H4* sentence data for all data measurement points according to ethnicity. 165
Figure 4.34: Boxplots displaying the values for the H1*−A1* sentence data for each of the auditorily identified phonation types for all data measurement points. 168
Figure 4.35: Boxplot comparison of the values for black and white speakers for the H1*−A1* interview data. 169
Figure 4.36: Boxplots representing the values for H1*−A1* sentence data for all data measurement points according to ethnicity. 170
Figure E19: Scatterplot of H1*−A1* data for the whole sample plotted against RMS energy. 287
Figure 4.37: Scatterplot of H1*−A1* data for the whole sample plotted against pF1 in Hertz (Pearson's *r*= 0.663). 171
Figure 4.38: Scatterplot of H1*−A1* data for black speakers plotted against pF1 in Hertz (Pearson's *r*= 0.658). 172
Figure 4.39: Scatterplot of H1*−A1* data for white speakers plotted against pF1 in Hertz (Pearson's *r*= 0.692). 172

Figure E1: Scatterplot of 2K*−5K data for the whole sample plotted against pF0. 278
Figure 4.40: Scatterplot of H1*−A1* data for the whole sample plotted against pF2 in Hertz (Pearson's *r*= −0.378). 173
Figure 4.41: Scatterplot of H1*−A1* data for black speakers plotted against pF2 in Hertz (Pearson's *r*= −0.388). 174
Figure 4.42: Scatterplot of H1*−A1* data for white speakers plotted against pF2 in Hertz (Pearson's *r*= −0.376). 174
Figure 4.43: Boxplots displaying the values for the H1*−A2* sentence data for each of the auditorily identified phonation types for all data measurement points. 176
Figure 4.44: Boxplots representing the values for black and white speakers for H1*−A2* for the interview data. ..177
Figure 4.45: Boxplots representing the values for the H1*−A2* sentence data for all data measurement points according to ethnicity. 178
Figure E2: Scatterplot of the 2K*−5K data plotted against pF0 for the black speakers. 278
Figure 4.48: Boxplots representing the values for H1*−A3* sentence data for all data measurement points according to ethnicity. 183
Figure E3: Scatterplot of the 2K*−5K data plotted against pF0 for the white speakers. 279
Figure 4.6: Scatterplot of the 2K*−5K data for the whole sample plotted against RMS (root mean square) energy (Pearson's *r*= −0.069). 135
Figure 4.7: Scatterplot of the 2K*−5K data for black speakers plotted against RMS energy (Pearson's *r*= −0.103). 136
Figure 5.1: Boxplots displaying the values for the SHR sentence data for each of the auditorily identified phonation types for all data measurement points. 188
Figure 5.4: Boxplots displaying the values for the CPP sentence data for each of the auditorily identified phonation types for all data measurement points. 192
Figure 5.5: Boxplots representing the values for black and whites speakers for the CPP interview data. 193
Figure 5.6: Boxplots representing the values for CPP sentence data for all data measurement points according to ethnicity. 193
Figure E31: Scatterplot of CPP data for the whole sample plotted against pF0 in Hertz. 293
Figure E32: Scatterplot of CPP data for black speakers plotted against pF0 in Hertz. 293
Figure E33: Scatterplot of CPP data for white speakers plotted against pF0 in Hertz. 294
Figure 5.7: Scatterplot of CPP data for the whole sample plotted against pF2 in Hertz (Pearson's *r*= −0.12). 196
Figure 5.8: Scatterplot of CPP data for black speakers plotted against pF2 in Hertz (Pearson's *r*= −0.153). 197
Figure 5.9: Scatterplot of CPP data for white speakers plotted against pF2 in Hertz (Pearson's *r*= −0.097). 197
Figure E34: Scatterplot of CPP data for the whole sample plotted against duration in milliseconds. 294
Figure 5.2: Boxplots representing the values for black and white speakers for the SHR interview data. 189
Figure E35: Scatterplot of CPP data for black speakers plotted against duration in milliseconds 295
Figure E36: Scatterplot of CPP data for white speakers plotted against duration in milliseconds. 295
Figure 5.10: Boxplots representing the values of black and white speakers for pF1 measured in Hertz. 199
Figure 5.11: Scatterplot of CPP data for the whole sample plotted against pF1 in Hertz (Pearson's *r*= 0.283). 200
Figure 5.23: Boxplots displaying the values for the HNR25 sentence data for each of the auditorily identified phonation types for all data measurement points. 214
Figure 5.12: Scatterplot of CPP data for black speakers plotted against pF1 in Hertz (Pearson's *r*= 0.219). 200
Figure 5.13: Scatterplot of CPP data for white speakers plotted against pF1 in Hertz (Pearson's *r*= 0.296). 201
Figure 5.14: Scatterplot of CPP data for the whole sample plotted against RMS energy (Pearson's *r*= 0.202). 202
Figure 5.15: Scatterplot of CPP data for black speakers plotted against RMS energy (Pearson's *r*= 0.385). 203
Figure 5.16: Scatterplot of CPP data for white speakers plotted against RMS energy (Pearson's *r*= 0.157). 203
Figure 5.17: Boxplots displaying the values for the HNR05 sentence data for each of the auditorily identified phonation types for all data measurement points. 206

Figure 5.3: Boxplots representing the distributions for SHR sentence data for all data measurement points according to ethnicity. 190
Figure 5.18: Boxplots representing the values of black and white speakers for the HNR05 interview data. 207
Figure 5.19: Boxplots representing the values for the HNR05 sentence data for all data measurement points according to ethnicity. 207
Figure 5.20: Boxplots displaying the values for the HNR15 sentence data for each of the auditorily identified phonation types for all data measurement points. 210
Figure 5.21: Boxplots representing the values of black and white speakers for the HNR15 interview data. 211
Figure 5.22: Boxplots representing the values for the HNR15 sentence data for all data measurement points according to ethnicity. 212
Figure 5.24: Boxplots representing the values of white and black speakers for the HNR25 interview data. 215
Figure 5.25: Boxplots representing the values for the HNR25 sentence data for all data measurement points according to ethnicity. 215
Figure 5.26: Boxplots displaying the values for the HNR35 sentence data for each of the auditorily identified phonation types for all data measurement points. 217
Figure 5.27: Boxplots representing the values for black and white speakers for the HNR35 interview data. 218
Figure 5.28: Boxplots representing the values for the HNR35 sentence data for all data measurement points according to ethnicity. 219
Figure E28: Scatterplot of SHR data for the whole sample plotted against RMS energy. 291
Figure E29: Scatterplot of SHR data for black speakers plotted against RMS energy. 292
Figure E30: Scatterplot of SHR data for white speakers plotted against RMS energy. 292
Figure 6.1: Example of nonconstricted creak phrase-finally for a male English speaker as provided in Keating, Garellek and Kreiman (2015). 228
Figure 6.2: A waveform display of a vowel extracted from the token *he* from the sentence data of speaker S3. 228
Figure 6.3: Waveform display of the word *ja* as spoken by speaker B2. 230
Figure 6.4: Waveform display of the word *ja* as spoken by speaker L2. 230
Figure 6.5: Boxplots representing the data distributions around the median for black and white speakers for CPP data for all data measurement points coded as modal. 240
Figure 6.6: Boxplots representing the data distributions around the median for black and white speakers for H1*−A1* data for all data measurement points coded as modal. 241
Figure E20: Scatterplot of H1*−A1* data for white speakers plotted against RMS energy. 287
Figure E21: Scatterplot of H1*−A1* data for black speakers plotted against RMS energy. 288
Figure E22: Scatterplot of H1*−A1* data for the whole sample plotted against pF0 in Hertz (Pearson's *r*= 0.164). 288
Figure E23: Scatterplot of H1*−A1* data for black speakers plotted against pF0 in Hertz (Pearson's *r*= 0.221). 289
Figure E24: Scatterplot of H1*−A1* data for white speakers plotted against pF0 in Hertz (Pearson's *r*= 0.152). 289
Figure E25: Scatterplot of SHR data for the whole sample plotted against pF0 in Hertz. 290
Figure E26: Scatterplot of SHR data for black speakers plotted against pF0 in Hertz. 290
Figure E27: Scatterplot of SHR data for white speakers plotted against pF0 in Hertz. 291

LIST OF TABLES

Table 3.1: Summary of measures used in this study and their relevance for voice quality……...103
Table 3.2: Coding categories and basic descriptions used as a guide in the auditory coding of the sentence data. 113
Table 4.1: The number of data measurement points coded according to auditorily identified phonation type for each of the ethnic groups for all of the read sentence data. 123
Table 4.2: Descriptive statistics for vowel duration for the sentence data between black and white speakers for different vowels (N=36;black=18, white=18). 129
Table 5.1: Statistical analysis results for the interview data.....221

LIST OF ABBREVIATIONS

2K(*)	The amplitude of the closest harmonic to 2000 Hz
2K(*)-5K	The amplitude of the closest harmonic to 5000 Hz subtracted from the amplitude of the closest harmonic to 2000 Hz
5K	The amplitude of the closest harmonic to 5000 Hz
A1(*)	The amplitude of the closest harmonic to the first formant
A2(*)	The amplitude of the closest harmonic to the second formant
A3(*)	The amplitude of the closest harmonic to the third formant
ANOVA	Analysis of variance
ARPA	Advanced Research Projects Agency
B	Used in this thesis as an abbreviation for auditorily identified breathy voice
C	Used in this thesis as an abbreviation for auditorily identified prototypical creak

CMU	Carnegie Mellon University
CPP	Cepstral peak prominence
CQ	Contact Quotient
Cw	Used in this thesis as an abbreviation for auditorily identified whispery creak
dB	Decibels
EGG	Electroglottograph
F	Used in this thesis as an abbreviation for auditorily identified vocal fry
f_0	Fundamental frequency
F1	First formant frequency
F2	Second formant frequency
F3	Third formant frequency
FFT	Fast Fourier Transform
Fw	Used in this thesis as an abbreviation for auditorily identified whispery vocal fry
GenSAE	General South African English
H	Used in this thesis as an abbreviation for auditorily identified harsh/aperiodic voice
H1(*)	The amplitude of the first harmonic
H1(*)−A1(*)	The amplitude of the closest harmonic to F1 subtracted from the amplitude of the first harmonic

H1(*)−A2(*)	The amplitude of the closest harmonic to F2 subtracted from the amplitude of the first harmonic
H1(*)−A3(*)	The amplitude of the closest harmonic to F3 subtracted from the amplitude of the first harmonic
H1(*)−H2(*)	The amplitude of the second harmonic subtracted from the amplitude of the first harmonic
H2(*)	The amplitude of the second harmonic
H2(*)−H4(*)	The amplitude of the fourth harmonic subtracted from the amplitude of the second harmonic
H2K(*)−H5K	An alternative way of expressing the measure 2K(*)−5K
H4(*)	The amplitude of the fourth harmonic
H4(*)−2K(*)	The amplitude of the closest harmonic to 2000 Hz subtracted from the amplitude of the fourth harmonic. “K” is a standard abbreviation for a thousand
H4(*)−H2K(*)	Same as H4(*)−2K(*) (see entry above)
HNR	Harmonics-to-noise ratio
HNR05	The harmonics-to-noise ratio between 0 Hz and 500 Hz
HNR15	The harmonics-to-noise ratio between 0 Hz and 1500 Hz
HNR25	The harmonics-to-noise ratio between 0 Hz and 2500 Hz
HNR35	The harmonics-to-noise ratio between 0 Hz and 3500 Hz
IP	Intonation phrase
K	Standard abbreviation for a thousand
L1	First language

L2	Second language
logEngergy	A variable used in this study: the logged values of RMS energy measure
logpF0	A variable used in this study: the logged values of pF0
logpF1	A variable used in this study: the logged values of pF1
logpF2	A variable used in this study: the logged values of pF2
M	Used in this thesis as an abbreviation for auditorily identified modal voice
OQ	Open quotient
PCA	Principal components analysis
pF0	Fundamental frequency estimated using the PRAAT algorithm
pF1	First formant frequency estimated using the PRAAT algorithm
pF2	Second formant frequency estimated using the PRAAT algorithm
PMII	Participant's Maori integration index
RMS	Root mean square
RP	Received pronunciation
SADV	Santa Ana del Valle
SAE	South African English
SHR	Subharmonics-to-harmonics ratio
SQ	Speed quotient

VPA	Vocal profile analysis
VS	VoiceSauce
W	Used in this thesis as an abbreviation for auditorily identified whisper

Chapter I: INTRODUCTION

1.1. BACKGROUND

Particular voice quality features shared by those speakers of a given language variety who belong to certain regional or social groups have been known to function as both indicators and markers (in the sociolinguistic sense) of those particular groups (Esling & Edmonson, 2011: 131). As noted by Esling & Edmonson (2011:131), since voice quality features are defined phonetically as the longest-term percepts of speech, they are also well-suited to conveying extra-linguistic meaning. Social meaning is one such form of extra-linguistic meaning and thus voice quality features are capable of functioning as sociolinguistic indicators or markers.

At least as early as the 19th-century, phoneticians have expressed that such long-term speech patterns could be used in the characterization of accent (Esling & Edmonson 2011:131). Sweet (1877) for example, notes how voice quality types may be characteristic of particular nationalities or individuals.

Sapir (1927:73) also expressed the view that a certain voice quality may index an individual speaker's social background, thereby reflecting the accent of a particular group, emphasizing that the voice comprises both social and individual components and that some facet of the voice has to be attributed to one's social background.

Abercrombie (1967:94) claims that there cannot be any doubt that specific voice qualities are able to function as recognizable characteristics of dialects or languages. This capacity of voice quality, says Abercrombie (1967:94) stems from those components of voice quality which are learnt. Where specific voice qualities do act as recognizable characteristics of language varieties, according to Abercrombie (1967:94), this indicates a predominance of these learnt

voice quality components over those which are unlearnt. Abercrombie (1967:94) cites the occurrence of 'adenoidal' voice quality in certain low-income urban areas, even for speakers without adenoids, who nevertheless adhere to the voice quality norms of their community, as one example of this[1].

Since it is clear that there is no physiological reason which can be used to account for such cases, it is profitable to consider other explanations which conceptualize voice quality phenomena as a product of the institutionalized norms of society, that is, as a form of learnt behaviour (Abercrombie 1967). The potential for voice quality features to function as differentiators of social groups, suggested here by Abercrombie (1967), will be taken up again in more detail in a separate section in the following chapter, dealing with the sociolinguistic function of voice quality variation.

Catford (1964) states that, in cases such as that discussed by Abercrombie (1967) as mentioned above, the function of phonation contrasts can be said to be non-phonological. Catford (1964:35) uses the term 'non-phonological function' to refer to the direct relation to the context which phonatory differences between speakers are able to convey, such that these differences may characterize both individual speakers, as well as dialects or languages and may thus index social variables such as social class and regional provenance.

Similarly, Laver (1968:50) states that a number of voice quality features may index social characteristics of speakers of a given accent, particularly those aspects of voice quality which, like other social behaviour, can be learned or imitated. This is the sense in which accents may be associated with certain voice quality features, such that those features may serve to index those social variables with which they are associated, for example, social status, attitudes, regional provenance and occupation (Laver 1968).

Ní Chasaide & Gobl (2010) also maintain that certain differences in voice quality have a sociolinguistic function to the extent that regional groups, social groups or linguistic groups may

[1] The link between income, health and 'adenoidal' voice quality is controversial and is discussed in more detail with reference to Knowles (1974) in chapter two, section 2.3.1. of this thesis.

display a tendency towards the usage of certain voice quality types. Ní Chasaide & Gobl (2010) further state that such features may also serve to signal certain social subgroups within a given dialect or language group.

In a recent sociophonetic study of one variety of South African English[2], Morreira (2012:126) observed impressionistically that there is a particular voice quality which may act as an indicator of the race of her informants. Morreira's (2012) sample consisted of Black middle-class speakers of English who attended ex-model C schools[3] for the majority of their schooling careers. Although the voice quality Morreira (2012) observes is reportedly more apparent for some speakers than for others, she claims that the presence or absence of this voice quality may indicate the race of a given speaker even if the speaker in question was identical to their White counterparts in terms of segmental pronunciation. Morreira (2012:126) also states that this observed difference in voice quality could potentially function 'as an identifying feature' in the absence of other phonetic cues to signal ethnicity. While this voice quality difference was not perceptible to a small group of students based on an informal perception experiment (suggesting that it may be below the level of consciousness), according to Morreira (2012), the difference in voice quality was evident to those linguists who had the opportunity to listen to the data.

In the current study, I aim to investigate, describe and evaluate the acoustic evidence for these observed differences in voice quality. I hypothesized that should certain voice quality features serve as linguistic indicators (and potentially, as markers) of ethnolinguistic identity[4] in the variety of English spoken by young, middle-class South Africans, these differences will at

[2] 'South African English' will be abbreviated to SAE henceforth.

[3] Morreira (2012:2) defines the term "…Model C Generation…" as being "…concerned with a category of former 'white' schools within the South African education system....," that is, those schools which were formerly reserved for White students. Morreira (2012) later explains that this term is used as a label for those who have either attended such schools or even those within the Black community who in their behaviour (both linguistic and non-linguistic) act as though they have attended such schools. Morreira's (2012) research participants were comprised of young self-identifying 'black' students who had received their school education from either private schools or from ex-model C schools.

[4] The term 'race' in this thesis refers to broader racial distinctions, for example, between black and white. 'Ethnicity' is used here to refer to sub-ethnic groups, such as for example, Xhosa. See Mesthrie (2016) for a recent and more detailed discussion of these terms as used in the sociolinguistic literature. In this study, the terms 'race' and 'ethnicity' are used interchangeably when referring to my sample, results and findings because, as will become clear when describing the sample characteristics, race and ethnic group overlap for this sample. In general, in this thesis I prefer the use of the term 'ethnolinguistic background' as a descriptor for categorization purposes to both 'race' and 'ethnicity' since it is both more accurate in describing the sample selection procedure as well as the nature of the differences observed given the current state of knowledge with regards to voice quality in South Africa.

the very least manifest themselves in the acoustic speech spectrum and would thus be observable. Such acoustic differences would be amenable to analysis and description by the researcher and would allow the researcher to provide inferences regarding the articulatory mechanisms involved in the production of the observed voice quality features, which could then be tested in future research, for example, using aerodynamic methods and ultrasound imaging. A supporting auditory analysis of such voice quality differences would also permit the researcher to more accurately describe what these differences may be.

Thus, although it will be necessary to discuss in some detail, the particulars of the auditory descriptive framework employed herein, it is not necessary to devote a large portion of this thesis to the evaluation of a particular descriptive framework (to the extent that this was done in Esling 1978, for example). Rather, the focus of the current study is on the evaluation of the usefulness and applicability of the relatively recently available acoustic methods in characterizing potential voice quality differences, should these be present, among speakers of General SAE[5] (GenSAE henceforth), in combination with an auditory analysis.

The third aim, which follows on from the abovementioned primary aim, is to identify which specific acoustic characteristics may be considered to constitute a voice quality indicator of ethnolinguistic identity (or potentially, a voice quality marker of ethnolinguistic identity) in GenSAE (although it would also be interesting to investigate other SAE varieties using these methods), to aid in the development of subsequent tests, including perception tests to pinpoint with greater accuracy, the specific articulatory and perceptual correlates involved. In addition, this research is expected to provide the fields of speech therapy, speech technology and telecommunications with a baseline of 'normal' voice quality data for GenSAE.

1.2. VOICE QUALITY THEORY: A BRIEF INTRODUCTION

Numerous writers in the past have commented on voice quality phenomena. A sizeable number of such historical works are thoroughly reviewed in Laver (1975). The intention of this section

[5] General South African English. This label is used here, as it is in Bekker (2009), to refer to the prestige (middle-class) variety of South African English, in theory at least, not restricted to any particular race group.

however, is to provide only a brief introduction to the work of some of the most prominent of these authors and their contributions towards developing a theory of voice quality which has been instrumental in shaping contemporary theory in this particular field of study. For a more comprehensive account of the history of voice quality research, which cannot be provided here due to space limitations, I encourage the reader to consult Laver (1975). A review of some of the more recent work in this field is provided by Podesva and Callier (2015) as well as by Garellek (2016).

Catford (1964)

The contribution of Catford (1964) to the field of voice quality research has shown itself to be invaluable, particularly since this author attempted to develop a comprehensive framework for the description of phonatory stricture types (such as "breath," "voice" and "whisper"). For each of these types specified, Catford (1964) provides critical and maximum airflow rates, the estimated percentage of glottal opening, an estimation of critical velocity, a description of the air-flow pattern involved, as well as the perceptual and acoustic effects of each type.

Abercrombie (1967)

For the specific laryngeal adjustments involved in producing differences in register, Abercrombie (1967:171) refers the reader to Catford (1964). Abercrombie (1967:101) also credits Catford (1964) for developing a base from which a precise phonation type nomenclature suitable for scientific description can be developed.

The framework offered by Abercrombie (1967:90) is not quite as elaborate as that of Catford (1964), but does make the distinction between 'voice quality' features and 'voice dynamics' features[6] and does not restrict his framework to the description of phonation alone.

[6] For Abercrombie (1967), voice quality features refer to those features contributing to general impressions of voice quality, while voice dynamics features result from the handling of the voice. See further section 1.4 on page 15.

Abercrombie (1967:93) uses the term 'voice quality' not only to refer to phonation, but also to sustained articulatory settings. Abercrombie (1967) derives the term 'articulatory setting' from Honikman (1964)[7]. Thus he provides 'adenoidal' voice quality, as an example, as that which results from a failure to relax velic closure (as a sustained muscular adjustment) and 'nasalized' voice quality as that resulting from an inadequate velic closure (Abercrombie 1967:93). As noted by Esling (1978:9), Abercrombie (1967:93) does not use the term 'articulatory setting' in the more restricted sense of Honikman (1964) where it applies only to settings which a language community shares.

In addition to voice qualities associated with such articulatory settings, Abercrombie (1967:93) also offers examples of voice qualities arising more directly from phonatory settings specifically, in other words, those which arise from the settings of muscular tension which directly influence vocal fold vibration.

Examples which Abercrombie (1967:93) provides in this regard include what he refers to as 'tight phonation,' characterized by an adjustment of the vocal cords allowing little air to escape and also, in contrast, 'breathy phonation' as that which involves the adjustment of the vocal cords allowing a lot of air to escape during vocal fold vibration.

Abercrombie (1967:100-101) also uses the term 'register' to refer to voice dynamic features of relatively short duration (although still of greater duration than segmental features),

[7] Honikman (1964:73) defines an articulatory setting in the following way: 'the disposition of the parts of the speech mechanism and their composite action, i.e. the just placing of the individual parts, severally and jointly, for articulation according to the phonetic substance of the language concerned. To put this another way, it is the over-all arrangement and maneuvering of the speech organs necessary for the facile accomplishment of natural utterance. Broadly, it is the fundamental groundwork which pervades and, to an extent, determines the phonetic character and specific timbre of a language' and as 'the gross oral posture and mechanics, both external and internal, requisite as a framework for the comfortable, economic, and fluent merging and integrating of the isolated sounds into that harmonious recognizable whole which constitutes the established pronunciation of a language.' Labov (1963:307) alludes to a similar concept in his Martha's Vineyard study, describing it variously as an 'articulatory style' and 'a favored articulatory posture.' Labov (1963:307-308) considers it 'a plausible mechanism for sociolinguistic interaction.'

which result from those phonatory modifications over which speakers have a certain degree of voluntary control, without being restricted to any particular pitch range.

Abercrombie (1967:101) provides ‘tight,’ ‘creaky’ and ‘breathy’ as illustrative examples of the impressionistic terms which are usually used to identify these different register types.

Laver (1968)

Laver (1968:45) acknowledges Catford’s (1964) contribution as being invaluable for the development of knowledge concerning phonation types by defining impressionistically labeled phonation types (such as ‘breathy voice’ and ‘creaky voice’ for example) in terms of vocal tract aerodynamics and physiology.

Laver (1968:45) does not consider pitch and loudness as ‘voice quality’ features proper but rather features of ‘voice dynamics,’ following Abercrombie’s (1967) distinction. However, since a number of the impressionistic labels used for the different phonation types often implicitly refer to pitch and loudness ranges which typically co-occur with those voice qualities, and because in the characterization of the speaker, such quasi-permanent features do appear to play a role, he does also refer to them in the model which he develops.

In his general phonetic approach, Laver (1968:46) focuses on the articulatory mechanisms involved in the production of a number of phonation types. For example, raising or lowering the larynx by modification of longitudinal tension is said to produce either ‘raised larynx voice’ or ‘lowered larynx voice’ respectively.

Laver (1968:47) thus sets out a labelling system for voice phenomena which does not make use of impressionistic descriptions based on single terms (for example ‘sepulchral voice’), as was the case for most previous work. Laver (1968:47) instead chooses to develop and use composite labels which are composed of phonetic terms for which physiologically-meaningful components are specified and has a particular focus on voluntary muscular settings as opposed to the physiological limitations (hence involuntary) underlying such settings. The system also

allows for the incorporation of scalar quantities indicating the degrees to which the vocal tract undergoes modification, as mentioned above (Laver 1968:48).

Laver (1975)

Laver (1975), in addition to providing an extensive historical review of the development of voice quality theory, provides us with a very important distinction (and one which will be employed in the current study in defining the aspects of voice quality of interest), namely that between 'extrinsic' and 'intrinsic' voice features. 'Intrinsic features' are defined as those features which 'lie outside the control of the individual speaker,' while 'extrinsic features' are defined as being constituted by 'all such choices of vocal activity, over which' a given speaker 'can exercise any degree of volitional control' (Laver 1975:26-27).

Thus intrinsic features cannot be learned and are not specific to a given culture, while extrinsic features are 'almost entirely social and psychological indices,' since they can be learned (Laver 1975:27). It is therefore the extrinsic features, using Laver's (1975:27) terminology, in particular, 'all the potentially controllable habitual muscular settings which characterize the manipulable component' of a speaker's voice quality, which are of primary interest in the current research, since it is these features which can be indexical of ethnolinguistic differences in voice quality.

Laver (1980)

Laver (1980) is perhaps one of the most comprehensive works on voice quality to date in which a classificatory framework (wherein laryngeal and other articulatory adjustments are specified) is presented.

Laver (1980:43) observes that a wide variety of labeling practices have been used for different voice qualities, but that with the exception of some systems used in speech pathology which are described as offering a certain level of phonetic precision, most other antecedent systems made use of only vague impressionistic labels. Laver (1980:43-44) favours a labeling system based on general phonetic theory, which does not aim at providing general labels, but

rather at the labelling of each of the independent physiological components involved in the production of each composite articulation.

Laver (1980:45) claims that there are sufficiently detailed descriptions of the physiological mechanisms involved in the production of certain phonation types including 'normal,' 'breathy,' whispery,' 'creaky,' 'ventricular' as well as 'harsh' voice, such that he chooses to focus on these types in developing his general phonetic descriptive framework.

Laver (1980) subsequently details the physiological adjustments involved in the articulatory settings which he describes and also specifies their acoustic effects as stated in the literature, providing the results based on an analysis of the recordings of Laver's own production of the various articulatory settings conducted by Francis Nolan. Thus Laver (1980) does not restrict his descriptive framework of voice quality to phonatory settings alone, but also incorporates and describes laryngeal, velopharyngeal, labial, faucal, pharyngeal, mandibular and lingual settings, as well as tension settings.

Laver (1980) discusses how the term 'register' has been used in the prior literature on voice quality. Laver (1980:93) does not include minor 'registers' in his outline of phonatory settings since the distinctions between them are too finely detailed and too numerous to be related directly to auditory percepts and specific articulatory modifications. Furthermore, the term is used ambiguously (Laver 1980). While usually covering modes of vocal fold vibration, the term 'register' may be used to also refer to pitch level, with different writers not always clarifying which meaning is intended (Laver 1980:93).

When the term is used by linguists, a different sort of ambiguity is involved. While the original technical term, which according to Laver (1980:93-94) was introduced to the field of phonetics by Henderson (1951), referred to the activity of the larynx, thus distinguishing it from supralaryngeal articulations, the term has since been used more as a phonological concept than as a phonetic one, where its use is extended to include supralaryngeal aspects. These include, for example, vowel articulation and tenseness and laxness of the speech organs (Laver 1980:93-94).

Laver (1980) thus concludes that at least for his purposes, the linguistic concept of 'register' is not a particularly useful one.

Laver (1980:94-95) subsequently provides descriptions for a number of phonatory settings, thus settings which apply specifically to how the vocal folds are used during phonation. He begins by providing a definition of 'modal phonation' as a neutral phonation mode in the production of which, only the true vocal folds vibrate efficiently and periodically, without audible friction. Laver (1980:94) then uses this definition in the description of non-modal phonation types in contradistinction to it, as those types of phonation where one or more of the aforementioned characteristics are not present.

In addition, Laver (1980:95) stresses that the phonation types defined in these terms are not wholly adequate and are merely descriptions, since the settings mentioned entail a great deal of complexity and are also greater in number than the foregoing descriptions would imply. Likewise, Laver (1980:95) considers 'modal voice' itself, when described in these terms as being merely a rough description rather than a comprehensive definition. Since coming up with a satisfactory and concise definition for 'modal voice' is problematic, Laver (1980:95) instead opts for an elaboration of its aerodynamic and physiological characteristics and these are then used as the basis for a discussion of the other types of phonation which he deals with.

1.3. MORE RECENT DEVELOPMENTS IN THE STUDY OF VOICE QUALITY

Having outlined the development of voice quality theory with reference to some of the more significant contributors in the development of current voice quality theory, the following section deals with more recent work investigating voice quality phenomena and those works dealing with voice quality theory, as well as other recent advances, such as the development of new technologies and methods which allow for more sophisticated and detailed analyses of voice source variation than was possible before. This section relies primarily on Esling and Edmonson (2011) who have provided a comprehensive summary of such recent developments. In section 2.9 on page 85, I discuss a more recent model of voice quality, the psychoacoustic model, where it is more relevant.

As noted by Esling & Edmonson (2011), the possible articulations of the pharynx have since been elucidated with greater precision than was possible before, which has provided an explanation for how the perceptual effects of lowered larynx voice relates to raised larynx voice, how lax voice relates to tense voice and how the effect of faucalization relates to pharyngealization through the possible range of adjustments of the pharynx as an articulator.

Esling & Edmonson (2011:134) advocate an integration of articulatory, acoustic as well as auditory knowledge in the formation of a theory which would allow for the meaningful interpretation of data derived by using instrumental methods. Esling & Edmonson (2011:135) emphasize the key role that the larynx plays in the production of sound. Spectral resonances are thus said to be significantly influenced by the cavity of the epilaryngeal tube including the surrounding pharyngeal area (Esling & Edmonson 2011:135).

Due to the size, potential changes in height and volume, flexibility as well as the relative independence from the adjustments of the oral articulator, the laryngeal articulator may, according to Esling and Edmonson (2011:135), be considered to provide a substantial contribution towards voice quality as a long-term speech phenomenon. A whole set of voice quality types may be ascribed to the possible fine changes induced by the laryngeal articulator (including pharyngealized voice, raised and lowered larynx voice and faucalized voice).

Esling & Edmonson (2011:136) also stress the need for using a suitable analogy in order for the effects of postural setting actuators[8] which result in different voice qualities and phonation types to be conveyed. According to Esling & Edmonson (2011:136) the appropriate analogy/taxonomy ought to be based on the functioning of valves (as Esling & Harris 2005 and Edmonson & Esling 2006 have also proposed).

Previous research by Esling and others have provided the suggestion that the structures of the throat function like brass instrument valves and that this is a suitable analogy (Esling &

[8] By which Esling & Edmonson (2011) appear to refer to the laryngeal structures which by means of their actions in concert give rise to certain laryngeal settings.

Edmonson 2011:136). Thus Esling & Edmonson (2011:140) conceptualize the vocal tract as consisting of a number of glottal and supraglottal structures which function like valves to modify airflow.

Following this analogy, most of the supraglottal valve-like structures act to augment the effects of the glottal vocal folds (also referred to in this analogy as valve 1) for the generation of a greater differentiation of the glottal waveform, as airflow is cut off by constriction, while one of the supraglottal valve-like structures lengthen or shorten the resonator of the pharyngeal tract (Esling & Edmonson 2011:141).

In addition to these theoretical refinements, Esling & Edmonson (2011:142) also point out that a number of acoustic measures have become available as potentially useful means of diagnosing types of voice quality, providing as one example, that of HNR (discussed in more detail in chapter 3).

Another fairly recent development, mentioned by Esling & Edmonson (2011:144) has been the development of correction algorithms for certain acoustic spectral tilt measures in order to compensate for the effects of formant frequencies and their bandwidths, as well as the development of software which is capable of calculating these corrected measures. The relevant developments to be used in the current research will be described in detail in the following two chapters (chapters 2 and 3).

Esling & Edmonson (2011: 146) also provide suggestions for sociophonetic investigations of voice quality phenomena. These authors suggest that techniques which they have outlined (as mentioned above) may be carried out on individual sections of a speech chain, (following Denes and Pinson, 1993) and these may subsequently be combined in a series, thus expressing a particular voice quality feature over longer stretches of speech (Esling & Edmonson 2011:146).

Esling & Edmonson (2011:146) also point out how the measures mentioned above (for example, spectral tilt and HNR) may be applied to vowels and subsequently, through

extrapolation, serve to characterize longer-term speech phenomena, i.e. over more than only one syllable. This particular procedure and how it relates to the current research aims will be explained in greater detail in the methodology chapter. Even more recent models used in accounting for voice phenomena will be discussed in the following chapter.

To summarize, in this section, I have briefly reviewed the works of the main contributors to the development of contemporary voice quality theory.

Research investigating such phenomena (indeed all phenomena which can be described as phonetic) is to a large extent limited by the analytical techniques and technology available. Some of the more recent advances in technology and techniques for voice quality analysis have been briefly mentioned in the foregoing section.

1.4. VOICE QUALITY DEFINED AND OPERATIONALIZED

It is necessary at this point, to clarify exactly which kinds of phenomena, given the wide range of phenomena subsumed under the general rubric of 'voice quality' by different authors, will be the focus of this study.

As has already been pointed out, what we are able to observe is in part determined by the available technology and techniques at the disposal of the researcher. I adopt the view of Esling & Edmonson (2011), that in order to offer a complete description of voice quality, ideally, it would be necessary to combine auditory, acoustic and articulatory analysis methods.

Due to the current unavailability (of some of the necessary equipment due to its presently prohibitive financial cost and other practical considerations), I do not consider it to be feasible at this stage of the investigation to use articulatory analyses (especially because of the exploratory nature of the current research. However, I hope that this will be possible in future once the relevant auditory and acoustic correlates have been established, in order to provide further confirmation of the findings presented in this work).

However, the current study does endeavor to use both auditory and acoustic analyses to shed light on the possibility of the existence of ethnolinguistic differences in voice quality in GenSAE (and thus potentially SAE more generally). These types of analysis are both more economically and practically feasible at this stage and can potentially be more straightforwardly applied in the analysis of voice quality in future, especially in situations where large and expensive equipment is not available (which would describe most field work situations, for example).

Therefore, the voice quality phenomena which are the focus of this research are those which can be captured by means of a combination of acoustic and auditory analyses.

It is necessary to provide at this point, a precise definition of voice quality, such that phenomena of interest are adequately circumscribed for analytical purposes and thereby also to clarify the nature of the phenomena which are of principal interest in the current investigation and which phenomena will be explicitly excluded.

In this thesis, I explicitly adopt the definition of voice quality as presented by Laver (1968:44), who, following Abercrombie (1967), defines it as 'the quasi-permanent quality of a speaker's voice' which, according to Laver (1968:44) derives from two principal sources, namely the anatomy and physiology of a given speaker's vocal tract as well as the settings, as described by Honikman (1964), that is, 'long-term muscular adjustments' of both the supralaryngeal vocal tract, as well as the larynx.

While a few refinements have been made to this definition since, mostly in terms of the elaboration of the articulatory mechanisms involved (as mentioned above), this basic definition retains its usefulness for circumscribing voice quality phenomena, in particular, as they relate to the interests of the current study.

Several points are entailed by the adoption of this definition which bear mention here. Firstly, voice quality is 'quasi-permanent' and one of the sources of which voice quality is composed is that of 'long-term' settings. As Esling & Edmonson (2011:131) point out, voice

quality can be considered to be the most persistent and the most long-term and most habitual of the phonetic components characterizing the voice of an individual speaker. Thus other prosodic phenomena which are not as long-term (such as the mid-length cues of stress, tone, intonation[9] and rhythm, that is, 'voice dynamic' features, following Abercrombie 1967) as well as segmental phenomena are not of primary interest in this thesis and the methods and techniques chosen are not designed to investigate these features of voice dynamics.

Another point entailed by the definition adopted here, is that both habitual phonatory setting as well as habitual supralaryngeal setting contributions to overall voice quality will be considered in interpreting the results. However, because several of the measures available (discussed in full in the following chapter) are designed to control for the effect of the supralaryngeal vocal tract, the results apply primarily to phonation.

One aspect of voice quality which is apparent in the adoption of Laver's (1968) definition of voice quality above which is not of interest in the current study is that of the first of the two principal sources of voice quality, namely the vocal tract anatomy and physiology of speakers. This is because effects which stem from components which speakers have no control over cannot function indexically (in the sociolinguistic sense), although they are indexical of other variables, not of primary interest here (these are discussed in more detail in the following chapter).

The definition I adopt therefore includes all phonatory settings primarily, as well as supralaryngeal settings to some extent, which could contribute to both auditory impressions as well as acoustic cues. As the exact relation between the articulatory adjustments and acoustic cues as well as auditorily identifiable voice quality percepts is not always clear, the former can, in the current study, only be inferred from the latter and would ultimately need to be confirmed using other methods (such as articulatory analysis).

Since it could not be determined with any degree of certainty *a priori* which specific voice quality may be involved and given the very exploratory nature of this research, I do not

[9] As Esling and Edmonson (2011) note, "long-term" is a relative concept. Thus while intonation patterns may extend over longer stretches of speech than segmental features for example, they nevertheless come and go more rapidly in speech than voice quality features do.

aim to test the hypothesis that a difference in terms of a specific voice quality exists (for example, the hypothesis that 'breathy voice' is used more often by young Black females than by young White females) as such, but rather aim at using the currently available techniques to uncover as broad a range of possibilities as far as the hypothesized differences in voice quality features involved are concerned and in order to aid in the interpretation of the results.

The combined results derived from the auditory and acoustic methods employed when interpreted correctly are well designed to achieve this aim. This approach potentially avoids the potential pitfall of failing to observe important differences by placing too exclusive a focus on one particular quality, which, given the few isolated comments made regarding ethnic differences in voice quality in the SAE literature, is perhaps the safest choice.

The motivations for the specific form which the auditory and acoustic analyses take will be provided in full in the methodology chapter. In the following chapter, I discuss the linguistic, paralinguistic and sociolinguistic functions of voice quality contrasts in reference to the relevant literature and also describe the acoustic measures used in this study.

1.5. FORMAL PROBLEM STATEMENTS

Earlier in this chapter, I pointed out that Morreira (2012) provided an indication of possible ethnolinguistic voice quality differences in South African English (specifically in GenSAE or potentially, in a 'crossover variety'[10]). This is the only work in which such observations have been made. The other studies which make isolated comments on this topic are reviewed in the following chapter. Thus, if SAE generally can still be described as an under researched dialect

[10] The term 'crossover variety,' or 'crossover accent' (where accent features are the focus) are used by Mesthrie (2017) to refer to the variety spoken by Black speakers who have adopted the norms of what used to be exclusively White SAE. 'Crossover speakers' is used to refer to those who speak with such a 'crossover accent' (Mesthrie, 2017). These terms are also adopted in this sense in this thesis. This term is not used in this sense by Morreira (2012) herself, except in terms of a possible 'crossover' feature, but the accent of a number of the speakers included in her sample matches the description of a 'crossover accent' as here conceived. It should however be noted that the speakers included in the current research are also considered to be GenSAE speakers. It is at this point still unclear in SAE scholarship to what extent and in what ways 'crossover speakers' (by definition, Black, although there may be some suggestion of some White speakers shifting pronunciation to something other than traditional White SAE) differ from GenSAE speakers (still currently mostly White). This issue will be taken up again in more detail in the methodology section dealing with sample selection.

cluster (see Bekker 2009), the topic of voice quality variation in SAE is even more so and has indeed been overlooked. While this may partly be because the state of SAE scholarship in this particular area has not been advanced sufficiently as yet (see comments by Mesthrie, 2017:25), it should be clear from the discussion earlier in this chapter, that significant advances, both in terms of theory and in terms of the analytic techniques and potential applications of these techniques for sociophonetic research, have been made in the field of voice quality research, in particular, by researchers such as Esling (2006; 2010) and others, reviewed in the following chapter.

That voice quality variation may have been overlooked in previous studies of SAE may in fact be considered a serious omission. This is because, for example, formant structure (and hence the formant measurements frequently used to characterize vowel quality differences in much sociophonetic research) is directly affected by articulatory settings[11]. It is therefore not only desirable, but indeed essential that attention be paid to the role played by voice quality variation in accent studies generally and for SAE specifically (as the focus of this study) that a proper account of voice quality is provided if we are ever to satisfactorily advance our understanding of the complexities of sociophonetic variation in the South African context.

There are numerous potential benefits of proceeding with such research. As one example, processes which would have otherwise remained unexplained can be accounted for in a much more straightforward manner with reference to articulatory settings, as stated mostly clearly by Laver & Trudgill (1979) (this matter is addressed more thoroughly in the following chapter).

[11] Laver (1980) provides a detailed description of such changes in formant frequency values linked to several articulatory settings. For example, Laver (1980:27;55) mentions the raising of F2 as a consequence of larynx raising and 'palatalized voice' and the raising of F1 associated with 'pharyngealized voice.' Cross-linguistically, for example, Gordon & Ladefoged (2001:400) note the association between phonation types such as 'creaky voice' and the raising of F1 (due to larynx raising) and a lowering of both F1 and F2 associated with an increase in pharynx width. Mennen, Scobbie, de Leeuw, Schaeffler & Schaeffler (2010:29) also note how formant frequencies may potentially be used in providing information on language-specific articulatory settings. For example, F2 values can be expected to be greater when there is a fronted tongue body setting, while F1 values could be expected to be greater in association with an open jaw setting (Mennen et al, 2010:29).

In addition, any developments in the field of speech technology as well as speech pathology in South Africa will in future need to take account of voice quality variation should evidence that systematic and pervasive variation of this kind in SAE be found to differentiate social groups.

1.6. FORMAL STATEMENT OF RESEARCH OBJECTIVES

This thesis thus has two main objectives. Firstly, the isolation, description and analysis, using acoustic means and a supporting auditory analysis, of potential ethnolinguistic differences in voice quality in GenSAE/'crossover variety,' thus providing scholarship with replicable results for SAE for future work investigating such phenomena.

The second related, yet subsidiary objective is to test the hypothesis that there is a difference in voice quality between young, English-dominant Black and White SAE speakers of similar educational backgrounds even for those speakers who are very similar to one another in terms of their segmental pronunciation of vowels. I will describe the evidence for such a difference in both auditory and acoustic phonetic terms. In the final chapter of this thesis, having already addressed this second objective, I will also provide a discussion of hypotheses which could account for the research findings and thus provide clearer directions for future research.

1.7. THESIS STRUCTURE

This thesis follows a fairly straightforward structure. The following chapter consists of a review of the literature relating to the different functions of voice quality variation, as well as a review of the literature pertaining specifically to the sociolinguistic function of voice quality variation, as the function which is of particular relevance to the current study. Also included is a review of studies which investigate sociolinguistic voice quality variation in a number of communities as well as ethnic differences in voice quality specifically. In addition, I review literature which investigates voice quality phenomena in isiXhosa as well as the comments made about

ethnolinguistic voice quality differences in SAE in particular. I also provide a review of the relevant literature on the particular measures used.

In the third chapter I deal with the methodology to be used in the current study and I describe the implementation of the acoustic measures and techniques selected as well as the motivations for their use. I also reiterate some of the information presented in the literature review contained in the second chapter of this thesis for the sake of readability. Some of the more technical aspects of the auditory coding procedure used for the exploratory data set for example, are included here rather than in the literature review chapter for this reason. I also describe the sample and sampling procedure and data collection techniques in detail. In the fourth and fifth chapters, I present the results of the auditory, acoustic and statistical analyses. The sixth chapter consists of a discussion of the findings and their interpretation. In this chapter, I also present new hypotheses which could account for the observed acoustic differences between the two ethnolinguistic groups. This somewhat unorthodox structure is motivated by the exploratory nature of this study, where no hypotheses regarding the use of a particular voice quality (for example, that black speakers use breathy voice more than white speakers) could be formulated initially given that no such specific claims have previously been made. Instead, in this thesis I first test the hypothesis that there is a difference in voice quality and having found acoustic evidence of such a difference, provide a discussion of hypotheses which could account for the difference which can then be tested in future research using articulatory and perception data. For this reason, I do not provide a detailed discussion of those hypotheses (which were developed later, based on the evidence) in this chapter. Appropriate conclusions are drawn based on the discussion in the final chapter.

1.8. CONCLUSION

In conclusion, I have provided the background to this study, as well as a brief discussion and review of some of the theoretical considerations involved in voice quality variation research. In addition, I have outlined new developments in both theory and analysis techniques and have presented the formal problem statement, research objectives and thesis structure. The following chapter provides a review of the relevant literature.

CHAPTER II: BACKGROUND

2.1. INTRODUCTION

This chapter has three main aims. Firstly, to provide a discussion of important theoretical issues in voice quality research and studies of ethnic differences in voice quality in particular. Secondly, this chapter aims to demonstrate the range of the types of methodologies commonly employed in the investigation of voice quality variation in particular communities and the conclusions which can be reached by utilizing such methods, with a specific focus on acoustic studies of ethnic differences in voice quality. Finally, I review the relevant literature pertaining to voice quality variation in South Africa with a particular focus on South African English and describe the measures used in this study. The following section deals with the topic of the indexical function of voice quality and reviews the work of the most prominent authors in this regard.

2.2. VOICE QUALITY AS AN INDEX

Catford (1964: 35) was one of the earlier writers to provide detailed comments on the different ways in which voice quality variation (specifically in terms of phonation) may function indexically. With regard to their purely linguistic function, Catford (1964:35) describes how differences in phonation types may be used phonologically, para-phonologically and non-phonologically.

Catford (1964:35) uses the term ‘phonological function’ to refer to those phonatory differences which register lexical differences or alternatively, differences in grammatical form. Catford (1964:35) includes as examples of such phonological distinctions, the voiced/voiceless distinction in English, contrasting lexical items such as *fat* and *vat*, as well as grammatical categories, such as between the verb *house* and the noun *house.*

Regarding the phonological function of phonation differences specifically, Catford (1964:35) provides the example of some languages in the Nilotic group where there is a contrast

between 'normal' voice and ligamental voice[12]. Abercrombie (1967:101) claims that in a number of languages, 'register'[13] contrasts have a phonological role, namely maintaining phonological distinctions, or in Abercrombie's (1967:101) words, functioning as carriers of 'language-bearing patterns.'

Ní Chasaide and Gobl (2010:453) note that for a number of South African, Native American and South East Asian languages, consonant and vowel contrasts are maintained by differences in voice quality. A competent review of the ways in which languages use voice quality phonologically is provided by Gordon and Ladefoged (2001).

Catford (1964:35) uses the term 'paraphonological function' to signify the direct correlation between differences in context and differences in phonation, providing the example of the difference between the use of whisper and (modal) voice in English. Catford (1964:35) notes that such a use of voice quality can be said to be paraphonological in function since while not correlated with any linguistic difference in terms of form (such as that already mentioned above for the Nilotic languages), there is a correlation with certain situational contexts, for example (modal/normal) voicing as used in situations where the context is unmarked, whereas 'whisper' is typically used in a "conspiratorial" context. For both the phonological and the paraphonological function, the differences in phonation are said to be linguistically contrastive (Catford 1964: 35).

According to Abercrombie (1967:101), one of the functions of 'register' is essentially paralinguistic, namely as an affective index, signaling the speaker's emotional state and attitude, a function which the majority of the world's communities make use of to at least some degree. Abercrombie (1967:101) points out that one need not assume that the same registers are used to signal exactly the same affective states in all languages and among all cultures.

[12] Following Catford (1964:32) this phonation type involves the tight occlusion of the arytenoid cartilages resulting in the restriction of phonation to the ligamental portion of the glottis, producing a [ʕ]-like, 'sharper' and 'clearer' auditory quality.

[13] Abercrombie (1967) uses this term to refer to phonation types, such as 'creaky voice' and considers 'register' to be one of the seven 'voice dynamics' features which also include loudness, pitch fluctuation, tempo, rhythm and tessitura. The use of the term 'register' to describe these sorts of features is generally problematic for a number of reasons (as pointed out by Esling and Edmonson 2011, among others) and thus will be avoided for the most part in this thesis.

Ní Chasaide and Gobl (2010) also discuss the paralinguistic function of voice quality variation. In their discussion, Ní Chasaide and Gobl (2010) refer to such paralinguistic aspects as those which involve temporal shifts away from a speaker's habitual voice quality in order to signal the emotion and mood of the speaker or to signal the attitude to either the content of what is being said or the speaker's attitude towards the listener.

As noted by Ní Chasaide and Gobl (2010), the signaling of affect by means of voice quality, is both bound by convention and may also be voluntarily controlled to some extent because it can also be employed for the purposes of deceiving the interlocutor. The focus of studies investigating the association between certain voice quality correlates and affective states, according to Ní Chasaide and Gobl (2010) has been on measuring parameters such as that of f_0 (including mean value changes, range, variability and the type of f_0 contour) as well as measuring mean value changes in intensity. In cases where emotional extremes occur, such changes in voice quality are, most probably, involuntary effects of the emotional state in question bringing about changes in physiology and if this is the case, then these specific changes can presumably be considered to be universal and therefore extralinguistic as they would therefore not be part of any learned system of conventions (Ní Chasaide and Gobl 2010). Gobl (2003:37) notes the increase in average values of intensity and in f_0 variability for a number of strong, although relatively different emotions, although also notes that the findings do not clearly show how listeners manage to differentiate between these emotions. For the interested reader, Kreiman and Sidtis (2011) provide a comprehensive review of literature on this topic.

Laver (1968:49) discusses how voice quality may be indexical of 'biological information,' for variables such as 'age,' 'sex,' 'physique' and 'medical state.' Laver (1968:49) finds it useful to separate permanent medical states and more temporary states, for analytical purposes.

More permanent aspects in this regard, would include factors related to overall health, such as the crude correlation between whispery/breathy and soft voices (also known as phonaesthenia) and ill health (Laver 1968:49). Other more permanent aspects would be the

correlation between permanent anatomical abnormalities (such as cleft palates) and a certain voice quality (Laver 1968:49).

Laver (1968:49) also mentions quasi-permanent states of health evidenced by voice quality features, such as those derived from a local inflammation of the vocal organs, as one finds in cases of tonsillitis or pharyngitis. Similar examples of quasi-permanent states include those related to changes in hormonal balance, such as in the case of pregnancy (Laver 1968:49 citing Perelló 1962) as well as during states of sexual arousal, which bring about changes in the mucus membrane of the vocal folds as well as the mucus which lubricates the larynx. The result of such changes may often be a slight increase in 'harshness' as well as either 'breathy' or 'whispery voice' (Laver 1968:49).

Generally permanent, but also occasionally reversible hormonal states, such as those as a result of thyroid, pituitary gland or adrenal disease associated with the reduction of the functioning of the endocrine system (including changes linked to voice disorders, such as endocrine dysphonia), are also expected to entail voice quality changes (Laver 1968:49, citing Luchsinger and Arnold 1965:135-136).

Laver (1968:49) also provides examples of states which are initially temporary but which may gain a degree of permanence, such as those effects (primarily due to vocal fold damage) which may result from the abuse of intoxicating agents such as tobacco smoke and alcohol.

According to Laver (1968), voice quality may also function as an index of psychological information and is used in assessing personality. Listeners appear to be prepared to come to quite far-ranging conclusions regarding psychological characteristics based on voice quality. The correlation between voice quality and characteristics of personality, according to Laver (1968:50) has been experimentally tested by several researchers. Voice quality changes may also signal an individual's mental health, as indicated by an appreciable number of studies of the correlations between acoustic parameters (such as those of intensity and f_0) and psychiatric illnesses such as depression and schizophrenia (Gobl 2003:33).

Ní Chasaide and Gobl (2010) also note that while there may be a universal element to the use of some particular voice qualities for certain specific communicative functions, there are also many cases when their use in communication is determined by culture. One example which they provide in this regard is that of sustained 'creaky voice' often used to signal 'bored resignation' for certain English speakers, while essentially the same voice quality is used to signal complaint or commiseration in Tzeltal Mayan (Ní Chasaide and Gobl 2010:456, citing Laver 1980: 126). Ní Chasaide and Gobl (2010) also point out how the differences in voice quality commonly found between males, females and children are primarily a reflection of differences in vocal tract anatomy, but that these can be either reduced or enhanced, according to socio-cultural settings, citing the example of the adoption of phonatory modes typically associated with males by women working in a predominantly male working environment. For a detailed discussion of the cultural constraints on voice quality variation as well as relevant issues pertaining to the indexicality of voice quality in general, the reader is encouraged to consult the more recent review by Podesva and Callier (2015).

In comparison to the phonological and paraphonological functions of voice quality differences as defined by Catford (1964:35), the non-phonological function is said to be non-linguistically contrastive, serving rather to distinguish between speakers of different languages or dialects. Under this heading are included those differences in phonation which index social variables, such as age, sex, social class and regional provenance as well as other variables such as the speaker's health (Catford 1964:35).

Abercrombie (1967:89-90) claims that there are three strands of communication. These are 'segmental features,' 'features of voice quality' and 'features of voice dynamics' (Abercrombie 1967:89). According to Abercrombie (1967), all three may function to signal idiosyncratic as well as personal traits, as well as traits which characterize social groups and traits characteristic of humanity in general.

Abercrombie (1967:94) maintains that voice quality characteristics can be recognized as distinguishing dialects from one another as well as distinguishing languages from one another. Abercrombie's (1967:94) interpretation of this is that in such cases where these characteristics do

serve to distinguish different languages or dialects from one another, there is a predominance of those elements of voice quality which are learned over those which are unlearnt. Features of 'voice dynamics,' according to Abercrombie (1967:95), are under the control of the speaker, such that, by being imitated from person to person, they may come to characterize both individuals and groups.

Ní Chasaide and Gobl (2010) also provide a discussion of the different functions of voice quality, restricting this discussion mostly to phonation, but also including in their discussion, voice qualities such as lax and tense voice, which involve different degrees of general tension, rather than referring specifically to laryngeal activity alone. Ní Chasaide and Gobl (2010:452) mention what they call the 'sociolinguistic function' of voice source variation in differentiating linguistic, regional and social groups.

This latter function of voice quality is most relevant to the interests of this study and so in the following section, I discuss this function in greater detail with reference to the relevant literature on the topic and some of the more well-known studies which have investigated the role of voice quality variables as indicators and as markers of sociolinguistic information.

2.3. THE INVESTIGATION OF THE SOCIOLINGUISTIC FUNCTION OF VOICE QUALITY VARIATION

Sapir (1927: 73) was one of the earlier writers to claim that 'voice,' has both social and individual aspects because we are able to imitate other people's voices to a considerable extent.

Sapir (1927: 73) makes the comparison between voice and gesture, stating that something of the voice has to be attributed to a speaker's social background in the same way that something about the physical gestures one uses is also dependent on one's social background. In the same way as gestures are particular to certain societies, the voice is also (Sapir 1927: 73). Laryngeal adjustments are specifically mentioned by Sapir (1927: 73) as being learned in this way.

Abercrombie (1967:92) claims that there are certain voice quality components which are beyond a speaker's ability to control and therefore cannot be learned (as described in Chapter I and in the previous section, using Laver's 1975 terminology 'intrinsic' features). Since this is the case, Abercrombie (1967:92) reasons that linguistic patterns cannot therefore be carried by these components and it is also not possible for them to characterize how the spoken medium is utilized in individual languages. These components can only index speakers who share certain physical traits, such as women as opposed to men (Abercrombie 1967:92). While it is not possible for these components to index social group membership, according to Abercrombie (1967:92), they can be used to differentiate individuals from one another.

In cases where certain dialects or languages can be recognized by unique voice qualities, or a unique use of voice quality, Abercrombie (1967:94) claims that the learnt components predominate over the unlearnt components. One example which Abercrombie (1967:94) gives of this is the presence of what he calls 'adenoidal voice quality' among those who do not have adenoids, having learnt this behaviour in conforming with the community norm for adenoidal voice quality found in certain urban slums. Abercrombie (1967:94-95) claims that a raised velum and velarisation are the main components which are used in imitating this voice quality and suggests that the characteristic Liverpool accent (often referred to as 'Scouse') originates from such circumstances.

Abercrombie (1967:95) regards what he refers to as 'voice dynamics' features as those which, since they are under the control of the speaker and since they are thus readily imitated, are able to characterize both individuals and social groups. For example, Abercrombie (1967:99) observes that for native Tlingit speakers, tessitura[14] is lower in comparison to that of English speakers, thus showing how the 'voice dynamics' feature of tessitura can be used to characterize, in this case, a language. This is due to its nature as an institutionalized feature which can be imitated (Abercrombie 1967:99).

Regarding 'pitch fluctuation,' Abercrombie (1967:102) claims that there are fairly obvious differences between both dialects and languages, which are responsible for the

[14] Used by Abercrombie (1967) to refer to the pitch range during normal phonation.

perception shared by some, that people who speak a different language sing when speaking. It is therefore conceivable that pitch fluctuations may act as an index.

Laver (1968:50) describes the various types of groups which can be indexed by voice quality features. Laver (1968:50) notes how particular voice qualities may be associated with certain accents, with voice quality therefore acting in part, as a pointer to the most typical speakers of a given accent and in this way, may index status in society, attitudes, values, regional provenance, as well as occupation.

2.3.1. Voice Quality as an Index of Regional Provenance

A number of studies have investigated the ways in which voice quality may index regional dialects. For example, by describing the articulatory setting (characterized roughly as having to do with articulations in the front of the mouth) as well as voice quality (roughly, all activity beyond the back of the tongue) of Scouse, Knowles (1974:98-99) demonstrated how such features came to function as an index of regional identity.

Knowles (1974: 99) explicitly uses the term 'voice quality' to refer to the same phenomenon as is referred to by Crystal's (1969) term 'voice stereotypes,' which is a more sociolinguistic conception of voice quality, namely as the 'vocal basis' which is not linguistic per se, but rather serves to identify social and regional groups. Knowles (1974:102-103) describes some possible differences between Scouse and RP regarding the various articulatory settings, based on his auditory and visual observations of his own speech.

Knowles (1974:111) suggests several possible articulatory factors contributing to the popular conception of the so-called 'nasal twang' and the 'adenoidal' quality commonly ascribed to Scouse, including the raising of the back part of the tongue, narrowing of the entire faucal area accompanied by the lengthwise contraction and forward movement of the uvula as well as upward movements of the larynx and pharynx.

Knowles (1974:115) suggests there may also be some 'whispery voice' in Scouse based on his impression that when he speaks Scouse, breath becomes depleted sooner than is the case

when speaking RP. He also notes the idiosyncratic use of creak for low-pitched speech by some Scouse speakers (Knowles 1974:115).

Knowles (1974:115) appeals to the appreciable breath-flow increase found towards voicing termination as the factor responsible for a few impressions of Scouse, such as that final vowels end with either a voiceless approximant or [h], that there is final aspiration and devoicing of underlying voiced consonants and that there is pre-aspiration of voiceless consonants in final position.

Knowles (1974:118) supports the view, expressed by Abercrombie (1976:95) that environmental factors played a role in bringing about the 'adenoidal' quality of Scouse. According to this account, environmental factors such as those causing diseases of the respiratory tract were responsible for the impairment of the nasal resonances of a fairly large number of people in the community (Knowles, 1974:118). This effect then came to be treated as sociolinguistic information in that it became a norm by which a given social group could be identified (Knowles 1974:118).

As a result, even those who had no vocal tract abnormalities would necessarily have to learn a way of producing this 'adenoidal' quality by means of compensatory adjustments in the same way that speech segments, for instance, are learned (Knowles 1974:118).

Knowles (1974:119) concludes by saying that 'Scousers do not use adenoidal quality because they or anyone else have or have had respiratory trouble, but because it makes them sound like Scousers.'

2.3.2. Voice Quality as an Index of Social Class Stratification within Regional Dialects

Norwich English, as described by Trudgill (1974), provides one illustrative example of how voice quality may index societal status (in this case, the distinction between middle-class and working-class speakers) as well as regional provenance. Trudgill (1974:186) claims that, based

on his observations of Norwich English, the setting[15] of the Norwich working-class serves to distinguish them from both middle-class Norwich speakers and from rural East Anglian speakers and from speakers of other English varieties. Trudgill (1974:186) also claims that this setting can be used to distinguish middle-class speakers as originally having been from the working-class in terms of background, even in cases where segmental pronunciation would indicate otherwise.

In terms of phonation type specifically, Trudgill (1974:186) describes the tendency towards employing a phonatory quality which can be described as 'creaky voice' among the Norwich working class, but not among middle-class speakers from Norwich.

According to Trudgill (1974:190-191), it is possible that the feature of the greatest significance in differentiating linguistically between different Norwich English speakers is the voice quality which results from the use of particular articulatory settings. This is said to be the feature distinguishing middle-class from working-class speakers most clearly (Trudgill 1974: 191).

Esling's (1978) study demonstrated how by combining techniques from experimental phonetics, auditory analysis as well as sociolinguistics and applying them to a particular speech community (the speech community of Edinburgh in this case), it is possible to describe the distribution of features of voice quality which are used in differentiating social groups within that community and to identify relevant social indicators, specifically those which, as Laver (1972:198) maintained, index features such regional provenance, occupation, social status and social attitudes and values.

In a preliminary study, Esling (1978:78) observed a consistent and regular distribution between boys from working-class areas who have different voice quality characteristics to those from middle-class backgrounds and suggests that the voice quality setting 'type A' (hypothesized to involve either larynx-raising or pharyngeal or faucal constriction) may be a feature which belongs to the vernacular of the working-class.

[15] Trudgill (1974) appeals to Honikman's (1964:73) use of 'setting' defined as: "the over-all arrangement and manoeuvring of the speech organs necessary for the facile accomplishment of natural utterance. Broadly, it is the fundamental groundwork which pervades and, to an extent, determines the phonetic character and specific timbre of a language…" In this study, the focus is on voice quality as defined in chapter one rather than on articulatory setting, as there is considerable overlap between these terms.

Esling (1978:122) subsequently recorded interviews which included a reading passage for fifty male informants (32 adult and 18 juvenile) native to Edinburgh from two Edinburgh wards (Morningside and Pilton) chosen as representative of socially stratified divisions within the city.

Esling (1978:108) subsequently devised an index similar to that which Labov (1966:211-220) and Trudgill (1974: 35-41) used in order to calculate social class index scores for every individual informant based on the five indicators: 'father's occupation,' housing type, education, occupation and residential area.

Esling (1978:138) then divided all 32 of the adult males included in his sample into three groups according to their social index scores. These groups were then compared according to the different settings specified in Laver's (1975) labeling system for voice quality.

Esling's (1978:144) auditory analyses pointed to pharyngeal and faucal constriction judgements being associated with the group with the lowest social index scores, hence providing the suggestion that these features may be more prominent in the Edinburgh vernacular dialect.

Esling (1978:175-176) concluded that the most obvious correspondence was between phonation type and the social divisions, observing an increase in creakiness ratings for speakers on the higher end of the social scale (representing the Edinburgh middle-class) and a decrease towards the lower end. Labels which indicate whispery or harsh phonation were used more frequently as the social index scores of the individual speakers decreased and lower social index scores were also associated with an increase in judgements of pharyngeal and faucal constriction as well as raised larynx (Esling 1978:176).

Esling (1978:272) also conducted a laryngographic investigation, which illustrated the main findings of the auditory analysis, namely an increase in harshness (or possibly 'ventricular quality')[16] associated with a decrease in social index scores.

According to Esling (1978:309), since the samples taken are representative, the constellation of the abovementioned features comprise the contrastive articulatory settings

[16] Instrumental evidence suggesting this included a laryngographic waveform pattern consisting of arrests in impedance occurring after positive peaks manifesting as a more 'flat-topped' appearance of the peaks, which suggested a "…momentary delay as impedance begins to increase" (Esling 1978:259).

distinguishing social groups of the Edinburgh community. Thus, those features found to be most common for those with higher social index scores, can be taken as a reflection of the features making up a middle-class articulatory setting, while the features most commonly found for the group with the lowest social index scores are representative of the speech of the vernacular dialect of Edinburgh (Esling 1978:309).

Esling's (1978:309) study demonstrated that voice quality features show social differentiation in the same way as other features of accent do and that by using an appropriate methodology it is possible to describe such variation in a particular community.

Stuart-Smith (1999:213) used a modified version of a VPA (or Vocal Profile Analysis) protocol developed by Laver, Wirz, Mackenzie & Hiller (1981), Mackenzie Beck (1988) and Laver (1991: 268) in order to make several important findings regarding the sociolinguistic function of voice quality in Glasgow.

Stuart-Smith (1999:215) also found support for class-based differences in phonation in Glaswegian English. While the voice quality of working class speakers is best described as featuring a strong component of 'whispery voice,' possible tongue root retraction, backed and raised tongue body as well as a more open jaw setting, the voice quality of middle-class Glaswegian English speakers is marked by an absence of these features (Stuart-Smith 1999: 215).

Stuart-Smith (1999:219) concludes that her analysis demonstrates clearly, the importance of understanding the role played by voice quality features in aiding the development of integrated descriptions of accents (Stuart-Smith 1999: 221).

Henton and Bladon (1985: 222) used the measure H1−H2 (a measure of the relative strength of lower to higher harmonics in the power spectrum reviewed later in this chapter) in their investigation of the habitual use of breathy voice for two British accents, namely RP and Modified Northern[17].

[17] Henton and Bladon (1985:222) define the Modified Northern accent in the following way:

Henton and Bladon (1985: 223) included 36 RP speakers, 16 males and 20 females, with 12 females and 13 males making up the total of 25 Modified Northern speakers (Henton and Bladon 1985:223).

Henton and Bladon (1985:224) found that the average first harmonic amplitude values for female RP speakers were greater than those of the amplitude of the second harmonic, in the order of 3.3 to 8.4 dB[18]. The enhancement on average, for males however, was only between 0.16 and 0.98 dB (Henton and Bladon 1985:224).

2.3.3. Voice Quality and Gender

In addition to its role in indexing social variables such as social class and regional provenance, voice quality may also play a role in signaling distinctions between genders within certain speech communities. For example, Stuart-Smith (1999: 215) found that Glaswegian males displayed a greater degree of 'creaky voice' with a greater degree of nasalization in comparison to Glaswegian females who showed evidence of 'whispery voice' to a greater degree as well as less nasalization. Henton and Bladon (1985:224) found that there were highly significant differences between males and females using the H1−H2 measure, with women having values 55.5 dB greater than males. Henton and Bladon (1985:224) claimed based on their results that women are breathier than men (at least in RP and Modified Northern accents). More recent research, most notably Simpson (2009) has however questioned the validity of the H1−H2 measure for comparisons between males and females.

Other studies, have investigated the role of voice quality in indexing the interaction between gender and variables such as regional dialect as well as race. Henton and Bladon (1988), for example, using an auditory analysis to identify creak-containing syllables and using chi-squared tests for significance found that males use creak significantly more in comparison to females (for both RP and Modified Northern), concluding that creak may therefore be considered 'a robust marker of male speech.'

'the accent of speakers who grew up and were educated in or near Leeds, but who have substantially moved away for substantial periods (e.g. to go to university or work) and have thus modified their native features.'

[18] In other words, they found that the amplitude of H1 for these speakers was between 3.3dB (for the vowel /ɒ/) to as much as 8.4dB (for the vowel /æ/) higher than the amplitude value for H2 when the mean value of the results for this difference calculation was calculated across all of the female RP speakers in the sample for each of the four vowels (/ɒ/, /ʌ/, /ɑ/ and /æ/).

Henton and Bladon (1988:21) observe however, that Modified Northern male speakers produce noticeably more creak than male RP speakers, which they interpret as either the exaggeration of female to male differences by Modified Northern males or alternatively, as the de-emphasis of these differences on the part of RP males.

More recent work by Podesva (2013) examined the role of gender and the social meaning of the use of non-modal phonation types in the speech of white and African American Washington DC residents. The sample for the study consisted of 32 speakers divided according to race and further subdivided according to sex, with the age of participants ranging from 18 to 75 (Podesva 2013).

The speech data relating to discussion of the local community was extracted from sociolinguistic interviews and was divided into IPs (intonational phrases) and each syllable was auditorily coded for six phonation types, modal, creaky, falsetto, breathy, whispery and harsh voice (Podesva 2013:429). Spectrograms, pitch tracks and waveforms as displayed in PRAAT were used to support the choice of label for phonation type on acoustic grounds (Podesva 2013:430). Linguistic factors examined in the study included the presence or absence of constructed dialogue, IP length, as well as the distance from the start of the IP as well as the end thereof (Podesva 2013:430). Social factors examined in the study included speaker sex, race and age and mixed effects linear regression as well as binomial regressions were used to analyze the data statistically (Podesva 2013:430).

Podesva's (2013:431) principal findings were that female speakers used creaky voice with greater frequency in comparison to male speakers and that while female speakers also exhibited a greater use of falsetto in comparison to male speakers, that it is African American women who contribute the most towards the falsetto pattern. Podesva (2013:434) found that white females use whispery voice more than white male speakers do, but did not find a difference between African American females and African American males for this phonation type. White males were found to use whispery voice less than all other groups included in the study (Podesva 2013:435). Age was found not to be significant in terms of the use of creak, thus younger women in Washington DC do not exhibit a greater likelihood of using creak in

comparison to older women (Podesva 2013:437). Podesva (2013) further examined the contexts in which falsetto is used in order to shed light on the discourse function of falsetto and concluded that this phonation type is often used by the African American women in Washington DC as a linguistic 'form of resistance'.

Khan, Becker and Zimman (2016), using a number of the same acoustic measures also used in my study, investigated the use of creak in American English. Khan, Becker and Zimman (2016) acoustically analyzed the vowels of 5 transgendered men (selected due to the observation that this demographic is associated with greater creak use, as found by Zimman 2012, 2013) and also collected perceptual data based on creak ratings of the speech samples derived from 14 listeners. They found significant correlations between creak ratings and lower f_0 as well as between creak ratings and greater irregularity (as evidenced by lower values for the harmonics-to-noise ratio). They also found non-significant correlations between creak ratings and several measures of spectral slope (discussed later in this chapter) and weaker multiple pulsing cues (as indicated by subharmonics-to-harmonics ratio values) (Khan, Becker and Zimman 2016). However, these correlations were in unexpected directions, prompting Khan, Becker and Zimman (2016) to conclude that their findings were most consistent with the use of one specific subtype of creak, namely 'Slifka voice' (discussed in more detail in the final chapter of this thesis where it becomes more relevant).

2.4. STUDIES OF ETHNIC DIFFERENCES IN VOICE QUALITY

Given that this study investigates ethnolinguistic differences in voice quality using acoustic techniques specifically, in this section I review research which investigates voice quality features as indicators of ethnicity and ethnolinguistic identity (with a particular, although not exclusive focus on those studies utilizing acoustic techniques to do so). In so doing, I discuss several broader pertinent issues relevant to such studies.

2.4.1. Anatomical Differences vs Learned Behaviour

A perennial (and as yet, unresolved) issue which often arises in studies of ethnic differences in voice quality is the question of whether observed differences between groups represent physiological differences between those groups or whether they are as a result of learned behaviour. Abercrombie (1967), as discussed above was one of the earlier writers to discuss the distinction between voice quality effects as a result of learned behaviour as opposed to anatomical differences. While at present the general consensus appears to be that learned behaviour (such as dialectal differences) is a more important contributor, ultimately better science will be required in order to adequately address this question (Kreiman and Sidtis 2011).

The literature concerning hypothesized racial differences in the anatomical structure of the larynx and the putative effects of such differences on the voice have mostly focused on American English speakers and in particular, African-Americans and Caucasian Americans (Kreiman and Sidtis 2011:147). The reader is encouraged to consult Thomas and Reaser (2004) for a thorough review of relevant perceptual studies.

Kreiman and Sidtis (2011:147) note that, if it is hypothesized that there are differences which can be reliably detected between Caucasian and African-American speakers when controlling for the influence of differences in dialect, by implication there should be consistent physiological differences as well as differences in vocal anatomy between those groups (Kreiman and Sidtis 2011:147). If consistent differences were found in terms of physiology, the implication of this would be that the speech signal should consistently code membership of a particular racial group (Kreiman and Sidtis 2011:147). However, if any observed differences are as a result of learned behaviour, such as that of dialect, speakers may either project such a racial identity or not (Kreiman and Sidtis 2011:147) which could conceivably result in a probabilistic tendency towards the use of certain voice quality features. Kreiman and Sidtis (2011:147) note that the issue of whether apparent differences between race groups stem from anatomical differences between races or rather from learned behaviour is an issue which the voice literature has not adequately dealt with in general.

Firstly, there are relatively few studies which have provided physiological data, a number of studies used experiments which did not consistently separate effects arising from organic

causes from learned effects and in perceptual studies, effects resulting from response biases are often not separated consistently (Kreiman and Sidtis 2011:147).

2.4.2. Physiological Studies

Kreiman and Sidtis (2011:147) cite Boshoff's (1945) study (also cited in Walton and Orlikoff 1994) in which 23 larynges from white South Africans were compared with those of 102 black South Africans (14 female and 88 male). Several minor distinctions between these two groups of speakers were reported. For example, the average height of the cricoid lamina was found to be slightly greater on average for black South Africans (males in particular) in comparison to Caucasian South Africans and black South Africans were also found to have longer vocal ligaments in most cases, as well as greater thyroid cartilage flexibility (Kreiman and Sidtis 2011:147). However, as Kreiman and Sidtis (2011:148) point out, because no information about the physical size or age of the subjects was provided, it is impossible to tell whether Boshoff's (1945) findings are indicative of actual anatomical differences between the races or whether they indicate an error in sampling. A more recent study by Ajmani (1990) reported no differences in terms of the cartilaginous structure of the larynx of adult Nigerian subjects (Kreiman and Sidtis 2011:148). Kreiman and Sidtis (2011:148) advise that for both of these studies, caution must be exercised in their interpretation given the substantial variation in laryngeal anatomy present even within race groups.

Kreiman and Sidtis (2011:148) also mention studies by Xue and Hao (2006) as well as Xue, Hao and Mayo (2006) in which volume and the relative size of the vocal tract were compared for African-American, Chinese and Caucasian speakers who were all precisely matched for sex and race. In this study, racial differences were found to be sex-dependent, such that for example, Caucasian females displayed 19 to 21 percent greater pharynx volumes than those of Chinese and African-American females, whereas for males, Chinese speakers exhibit greater oral volumes overall than those observed for African-American and Caucasian males (Kreiman and Sidtis 2011:148).

2.4.3. Perception Studies

As Kreiman and Sidtis (2011:148) point out, most of the earlier studies using perception tests of speaker race identification (most of these having been investigations of American English) did not control for dialect, such that for many of these studies, it is not possible to establish that race rather than dialect contributed towards the observed differences. Kreiman and Sidtis (2011:148) cite Dickens and Sawyer (1962) as a prime example of such studies. Dickens and Sawyer (1962) recorded sentences spoken by ten African-American and ten Caucasian American speakers and played these samples to groups of Caucasian and African-American listeners (Kreiman and Sidtis 2011:148). They found that listeners were able to identify speaker race at levels above chance (Kreiman and Sidtis 2011:148). However, listeners also displayed a same-race 'response bias,' such that white listeners judged more of the speakers to be white and black listeners judged more speakers to be black (Kreiman and Sidtis 2011:149).

Similar findings were reported by Lass, Tecca, Mancuso and Black (1979) and Lass, Almerino, Jordan and Walsh (1980) (Kreiman and Sidtis 2011:149). For these studies, as Kreiman and Sidtis (2011:149) note, accuracy was found to decrease with decreasing stimulus length, such that for vowels, accuracy was at chance levels. This pattern of decreasing accuracy as articulatory information decreases would suggest that rather than being innate or anatomical, cues to race constitute a learned articulatory behaviour (Kreiman and Sidtis 2011:149).

Walton and Orlikoff (1994) conducted a perception study in which extracted samples of recorded sustained[19] vowels produced by 50 white and 50 black American English speakers were presented in the form of an ordering task in race-matched pairs to 12 listeners. Walton and Orlikoff (1994) found that listener accuracy in the task of race identification was significantly greater than would be expected by chance, however Kreiman and Sidtis (2011) criticize this study for not including any formal controls for the effect of dialect (which refers to the overall effect which dialect, in terms of both segmental and suprasegmental features, could be expected to have in a speech perception task of this nature).

[19] The terms 'continuous' and 'sustained' as employed here refer to the types of speech material used. 'Continuous vowels' are those extracted for measurement from continuous speech, whereas sustained vowels are vowels uttered in isolation, typically sustained over a number of seconds and thus usually longer than most vowels extracted from continuous speech.

Thomas and Reaser (2004) in contrast, did control for the effect of dialect, observing that both Caucasian and African-American listeners found it difficult to accurately recognize African-American speakers as African-American (approximately 56% correct recognition), but were considerably better at recognizing as Caucasian, Caucasian standard American English speakers (approximately 94% correct recognition) as well as standard African-American speakers as African-American (with approximately 92% correct).

Further afield, Szakay (2008:2) performed a perception test in order to investigate the role that suprasegmental cues play in ethnic dialect identification by naïve 'Pakeha English'[20] and 'Maori English' listeners in New Zealand, using seven distinct conditions to test listeners' abilities.

A couple of the conditions used by Szakay (2008:53) in this experiment were expected to retain some voice quality information. For example, one condition involved the removal of unusually high pitched tokens and would therefore be likely to preserve creaky voice tokens. The 'intonation only' condition (Szakay's 2008 condition 4) as well as the 400 Hz low-pass filtered stimuli would also be expected to retain some voice quality information.

Szakay (2008:75) found that a high level of accuracy (75%) was obtained by Maori listeners in identifying speakers of Maori English in the low-pass filter condition, suggesting that voice quality could be considered as a 'useful cue' in the task of ethnic dialect identification.

In a follow-up study investigating the influence of voice quality on ethnic dialect identification specifically, Szakay (2012:389) conducted a perception experiment using 15 second long sound files (sampled from interviews of 36 New Zealand English speakers, 24 Maori and 12 Pakeha), in which 52 Maori and 55 Pakeha listeners, all native New Zealanders (thus 107 in all) participated.

Szakay (2012:391) used logistic regression models in order to examine the sensitivity of listeners to the characteristics of voice quality which had been identified. These models included the measure of average (unnormalized) H1−H2 for each speaker as a possible independent

[20] 'Pakeha English' refers to 'the English spoken mainly by European New Zealanders' (Szakay 2008:9).

variable, while the dependent variable was the perceived ethnicity of the speaker (Szakay 2012:391).

The results revealed that average H1−H2 for a speaker significantly predicted the perception of ethnicity, interacting with the PMII (participant's Maori Integration Index-a measure of Maori English exposure and Maori community involvement) scores of listeners (Szakay 2012:391). Speakers with high values for H1−H2 (interpreted by Szakay 2012:393 as 'breathy') were perceived by native New Zealanders as Pakeha, while those with low values for this measure (thus 'creaky,' according to Szakay 2012:393) were perceived as Maori.

According to Szakay (2012:391), this finding revealed that the PMII score was a more important predictor of performance on the ethnic dialect identification task than the ethnicity of the listeners, since the effect of ethnicity did not significantly affect the participants' accuracy. This suggests, according to Szakay (2012:391), that the more important factor than ethnicity alone in dialect identification accuracy was that of the extent of Maori community group membership and involvement. This result, once again, suggests that learned, sociocultural factors are of greater importance than anatomical differences.

In Perrachione, Chiao, and Wong's (2010) study, one subset of African American voices were consistently misclassified as Caucasian by both Caucasian and African American listeners, using stimuli of sentence-length, but not the other way round (Kreiman and Sidtis 2011:149). Once the listeners received training to help recognize the misclassified subset of African American speakers, Caucasian listeners showed an advantage in recognizing speakers of their own race (Kreiman and Sidtis 2011:149). These results, according to Kreiman and Sidtis (2011:149) again point to dialect rather than anatomy as the source of apparent racial differences in voice.

Commenting on the results of this study, Kreiman and Sidtis (2011:150) conclude that because as articulatory information declines so does race classification accuracy, this would suggest that the most salient information is primarily contained in patterns of articulation, which would point to learned differences (for example, as part of dialect), rather than organic differences.

Ng, Chen and Chan (2012) offer an interesting perspective regarding acoustic studies of ethnolinguistic differences in voice quality specifically. They cite a study by Awan and Mueller (1996) for example, who found lower mean speaking fundamental frequencies for African American kindergarteners in comparison to their Hispanic counterparts and suggested that this f_0 difference could be attributed either to anatomical or linguistic differences. Xue, Needley, Hagstrom and Hao (2001) also found lower mean speaking f_0 for African American as opposed to Caucasian American speakers between the ages of 70 and 80 years (Ng, Chen and Chan 2012). Ng, Chen and Chan (2012) clearly state that in order to determine whether speaking f_0 differences arise from linguistic rather than interracial differences in anatomy, the effect of difference in language also needs to be investigated and they cite various studies which do this. For example, Andrianopoulos, Darrow and Chen (2001) found significantly greater mean values for f_0 for Mandarin-Chinese speakers (both male and female) in comparison to Hindi speakers from India and American English speakers (both Caucasian and African American) (Ng, Chen and Chan 2012). This study was however critiqued for its use of isolated vowels by Altenberg and Ferrand (2006).

Altenberg and Ferrand (2006) investigated the speaking f_0 of English and Cantonese using continuous speech data from 9 bilingual female speakers of the same ethnicity, but did not find any significant differences. Ng, Chen and Chan (2012) point out that for a number of the English/Cantonese bilingual speakers used in this study, it was likely that English was dominant with Cantonese as an L2 (participants rated their English proficiency as equal to or greater than their proficiency in Cantonese, claimed that competency and comfortability was greater for speaking English and reported using English more than half of the time) and suggest that different findings would be expected if language dominance was reversed. Ng, Hsueh and Leung (2010) found in comparing the f_0 range and mean f_0 for bilingual Cantonese-English preadolescents, that Cantonese has significantly lower values for these measures than English. The results, Ng, Chen and Chan (2012) assert, indicate that language background has an effect on f_0, even when the vocal apparatus is the same (given that they sampled the same bilingual speakers when speaking in two different languages). Ng, Chen and Chan (2012) point out that ethnicity and age (the latter constituting an intrinsic determinant of voice quality, the former at least hypothetically intrinsic using Laver's (1975) terminology) as well as language, a learned behaviour, may affect vocal characteristics. If the interest of a given study is primarily to

investigate vocal quality differences linked to language alone, it would thus be necessary to exclude the effects of both ethnicity and age (Ng, Chen and Chan 2012:e172). Ng, Chen and Chan (2012:e172) suggest that one way of doing this would be compare vocal characteristics based on utterances in two languages as spoken by bilingual speakers with roughly equivalent levels of proficiency in the two languages of interest.

2.5. ACOUSTIC STUDIES OF ETHNIC DIFFERENCES IN VOICE QUALITY

Since the interest in the current study is primarily focused on the acoustic measures, it is necessary to provide a review of those studies which have likewise used acoustic measures to investigate ethnic differences in voice quality. In this section, the focus is on those acoustic studies in particular which are closest in terms of research aims and methodology to the current study. As pointed out by Kreiman and Sidtis (2011:148), formulating hypotheses regarding which acoustic cues should be investigated in studies of hypothesized anatomically-linked differences in race is challenging since there is so little evidence in the form of physiological data on which to base such hypotheses.

Those studies which provide acoustic data have often found conflicting results. For example, Hollien and Malcik (1967) and Hudson and Holbrook (1982) reported lower overall fundamental frequency for African-American subjects. Walton and Orlikoff (1994) however reported no significant differences in mean f_0, while Morris (1997) and Xue and Fucci (2000) also reported no significant average f_0 differences, shimmer (cycle-to-cycle amplitude perturbation), jitter (cycle-to-cycle frequency perturbation) and spectral noise (Kreiman and Sidtis 2011:148).

Walton and Orlikoff (1994) however, found that Black speakers displayed significantly greater average jitter and significantly greater average shimmer than their White counterparts for correctly identified speaker pairs. For incorrectly identified voice pairs, Walton and Orlikoff (1994) found no significant differences in terms of the noise measures. Walton and Orlikoff (1994) also found a statistically significant difference in shimmer between the two groups of

speakers and a significantly lower average HNR (harmonics-to-noise ratio) for Black speakers in comparison to their White counterparts.

Walton and Orlikoff (1994) state that the results of their research support the assertion by Murray (1988) that laryngeal features have a prominent role to play in the task of speaker identification. In discussing their results however, Walton and Orlikoff (1994) note that:

'Although as much linguistic information as possible was removed from the samples used in the present study, it is difficult to know whether the greater noise in the Black voices was a function of physiological differences or whether it was due to some paralinguistic feature that is learned within the context of dialect.'

Indeed, one of the criticisms leveled by Kreiman and Sidtis (2011:149) aside from not controlling for response bias, is that Walton and Orlikoff (1994) did not formally match for the effects of dialect. Thus, while Walton and Orlikoff (1994) suggest, citing Boshoff (1945), that the acoustic differences observed may be as a result of racial differences in vocal fold physiology between black and white speakers, these authors nevertheless concede that, as stated by Dillard (1972), an expert on Black English Vernacular, in comparison to racial factors and geographic factors, social factors are greater in importance in the determination of dialect variation.

Walton and Orlikoff (1994), as well as Andrianopoulos, Darrow and Chen (2001) did not find any systematic variation in formant frequencies according to race and following Sapienza (1997) glottal airflow patterns also do not show such systematic variation (Kreiman and Sidtis 2011:148).

Newman and Wu (2011:163) conducted a study in which they utilized acoustic measures to indicate voice quality differences between American English speakers in a sample consisting of White, Black, Latino, Korean and Chinese Americans, including shimmer, jitter and the unnormalized spectral tilt measure, H1−H2 (following Szakay 2008).

While the measures of shimmer and jitter did not show many significant effects (except for a few on the individual level), phonation type (as indicated by the measurements of the spectral tilt measure H1−H2) showed a clear difference between most non-Asians and Asians (Newman & Wu 2011:165-166). Higher values for spectral tilt were found for the two Korean

females and all of the Asian American males included in the sample, in comparison to the non-Asians (Newman & Wu 2011:166). The data, according to Newman & Wu (2011:166), support the 'idea' that a component of 'sounding Asian' is a comparatively breathy voice.

Ng, Chen and Chan (2012) analyzed recordings of 40 bilingual Cantonese-English speakers (20 female and 20 male) reading passages of text in both Cantonese and English by means of long-term average spectra (LTAS) and f_0 in Praat. From these measurements, they derived mean speaking f_0, and quantified three LTAS parameters, namely first spectral peak (theoretically indicative of the level of vocal fold stiffness), mean spectral energy (a correlate of laryngeal tension) and spectral tilt (with low spectral tilt being associated with vocal fold hyperadduction) (Ng, Chen and Chan 2012). They found significantly lower f_0 values for Cantonese in comparison to English when spoken by female speakers, who also exhibited significantly higher first spectral peaks than males for both English and Cantonese (Ng, Chen and Chan 2012).

For all speakers (both male and female) spectral tilt was significantly lower and mean spectral energy significantly higher when speaking Cantonese as opposed to English (Ng, Chen and Chan 2012). Ng, Chen and Chan (2012) were thus able to hypothesize that when speaking English, female bilingual speakers use a higher rate of vocal fold vibration than when speaking Cantonese and that bilingual speakers of both sexes used greater laryngeal tension as well as more aspiration noise and a breathier voice quality when speaking Cantonese.

They conclude from their study that even when the same vocal apparatus is used, language itself has an effect on a speaker's vocal quality, which can be associated with the use of different phonation types (Ng, Chen and Chan 2012:e175).

Szakay (2012:384) investigated apparent voice quality differences for the ethnic dialects of New Zealand English using H1−H2. This study examined the speech of 36 speakers (24 Maori and 12 Pakeha), equally matched for speaker sex and ranging in age from 18 to 65 years (Szakay, 2012:386).

Szakay (2012:386) used PRAAT to manually segment the consonants and vowels and thereafter, H1−H2 values for all vowels (approximately 40 for each speaker) were calculated

using a PRAAT script which took a measurement of the amplitude of the greatest peak in amplitude within ten percent of f_0 and the peak which was twice f_0 and then calculated the difference between these two maxima.

A linear regression model was fitted to the data, revealing a significant interaction between three variables, namely gender, ethnicity and age predicting H1−H2 values (Szakay 2012:387).

Szakay (2012:393) suggests that follow-up studies are needed which evaluate other phonation types, including hoarseness and harshness as well as studies which use other measures such as cepstral peak prominence to investigate these phenomena further.

Szakay and Torgersen (2015) investigated ethnic, gender and regional differences in voice quality for London English using the acoustic measures H1−H2 and f_0 derived by means of a Praat script. They compared 28 speakers (9 female and 18 male) from Hackney, an area representing Inner City London, with 14 speakers from Havering (equally balanced for gender), representing Outer London. In terms of ethnicity, the entire Havering sample consisted of Anglo speakers, while the Hackney sample consisted of 9 Anglo speakers and 19 non-Anglo speakers (descendants of immigrants), intended to represent the ethnic composition of the two respective areas. Szakay and Torgersen (2015) found that the voice quality of Hackney speakers was significantly breathier than that of Havering speakers (attributed mostly to the greater breathiness observed for Hackney males from both ethnic groups), male non-Anglo Hackney speakers had significantly lower pitch in comparison to both Hackney Anglo speakers and male Havering speakers, and female Anglo Hackney speakers had significantly lower pitch than their non-Anglo counterparts.

Szakay and Torgersen (2015) further examined the voice quality effects for the Hackney data (only this data set allowed for an analysis of the effect of ethnicity) using mixed effects linear regression. They found that vowels with greater intensity were creakier, lower vowels were slightly breathier and that for male speakers, longer vowels were slightly breathier (Szakay and Torgersen 2015). Ethnicity, f_0 and gender were also found to have a significant interaction effect on H1−H2 values such that higher pitch corresponded with higher H1−H2 values only for female Anglo speakers, whereas there was no significant effect for f_0 for female non-Anglo

speakers. Male speakers from both locations, in contrast, showed a negative correlation between f_0 and H1−H2.

Since Hackney males display higher H1−H2 values than Hackney females, in spite of the traditionally held view that females are breathier than males according to this measure and that if anything, the H1−H2 tends to overestimate this difference, Szakay and Torgersen (2015) reason that this constitutes evidence that voice quality variation is utilized for sociolinguistic (i.e. potentially indexical) purposes and that both pitch and voice quality should be considered as "innovative features" characterizing the speech of inner city London.

The foregoing review has described a number of key studies which have examined voice quality differences as they relate to ethnic dialects in general. In so doing, a number of methods which are available for investigating such phenomena have been mentioned, some of which will be applied directly in the current investigation and some others will be applied in a modified form, as will be fully explained in the following chapter. The following section of the current chapter reviews the literature specifically pertaining to voice quality research in South Africa, with a focus on SAE in particular.

2.6. VOICE QUALITY VARIATION IN SOUTH AFRICAN ENGLISH (SAE): A REVIEW OF THE RELEVANT LITERATURE

Apart from a few comments offered by Bekker (2009) which are discussed below, there appears to be no mention in the literature of any specific voice quality features which clearly serve to distinguish GenSAE[21] from varieties of English such as RP.

Bekker (2009:130) observes that for his GenSAE subjects, there is a tendency to use creak during the latter parts of vowel production, although he later notes that it is a general feature of vowel pronunciation (Bekker 2009:432). This 'creak' is said to be accompanied in many cases, by what Bekker (2009:432) describes as 'pre-aspiration' before the occlusion at the

[21] General South African English, the prestige, middle-class variety of South African English.

end of each word, specifically before voiceless oral stops (the example provided by Bekker 2009:432 is of the word *bat*). Bekker (2009:130) explains in a footnote what he means in this case, by the term 'pre-aspiration,' using the term essentially as it is used by Foulkes and Docherty (2006).

As Bekker (2009:130) explains, these authors use the term 'pre-aspiration' to refer to two principal patterns. The first of these is an extended form of frication, for which there is an increased concentration of fricative energy in the higher frequency region preceding stop gaps (Foulkes and Docherty 2006:417-418).

The other form of pre-aspiration described in Foulkes and Docherty (2006), contains a more even spread of energy, which is said to be an indication of aspiration as opposed to frication. Bekker (2009:130) links this latter aspiration phenomenon to the phenomenon of 'pre-aspiration' that he observes in the speech of his informants.

Bekker (2009:130) also cites research which indicates that this particular feature of 'pre-aspiration' can be employed indexically since it is within the speaker's control. In this regard, Bekker (2009:130) cites Foulkes and Docherty's (2006) finding of gender stratification in terms of increased use of pre-aspiration as described above in the speech of (in particular, young) Newcastle females.

As further evidence that this is a feature which is under the control of the speaker and therefore may potentially be indexically employed, Bekker (2009:130) cites Foulkes and Docherty (2006), who find that such pre-aspiration is absent from an equivalent set of data drawn from a sample of Derby speech.

Bekker (2009:130) suggests, based on Lieberman and Blumstein (1988:59), that both this phenomenon of pre-aspiration and that of creak observed for his subjects can be explained as a result of the opening of the vocal chords as phonation ends (and additionally lowering air pressure from the lungs).

Bekker (2009:432; 130) suggests that both the observed creak and pre-aspiration may be prestige features in GenSAE and concludes that the possible indexical use of this observed pre-aspiration phenomenon in SAE is worth further investigation.

The only writer known to me to comment unequivocally on the possibility of ethnic differences in voice quality in SAE is Morreira (2012). Morreira's (2012) research was not aimed at uncovering voice quality differences, however it does offer explicit comment on the possibility of ethnic differences in voice quality in SAE. Morreira (2012:2) characterizes her research as an attempt at investigating SAE since the new socioeconomic and sociopolitical changes which have taken place in South Africa, with the particular focus of her study being that of the English accents of 'the new black middle class youth.'

Morreira's (2012:3) work assumes that the English accents of this particular group will be more similar to those of White SAE (but perhaps not identical to them) than to those of traditional Black SAE.

Morreira (2012:20) provides a political and social history of education in South Africa and analyzes data taken from interviews of 44 University of Cape Town students who were self-identified as 'black.' The participants also read out a word list to allow for an analysis of their more formal speech style. Both the interview and word list data were analyzed, with a focus on the acoustic analysis of the GOOSE vowel (of Wells' 1982 lexical sets) in particular (Morreira 2012:20).

Morreira (2012:127) states that impressionistically, either intonation pattern, or timing, or both as well as voice quality clearly affect her informants' speech. While not offering an explanation of what this voice quality might be, Morreira (2012:126) does assert that it would indicate the 'race' of the speakers even should the speakers be segmentally identical to White SAE speakers. Morreira (2012:126) also notes that this particular feature was observed to be more prominent for some speakers than for others. Morreira (2012:127) intentionally does not

attempt to offer speculation regarding whether this is a conscious effect or not, but says that it is noticeable and deserves further research.

Morreira (2012:127) evidently conducted an informal test with UCT students to help ascertain whether they were able to recognize this voice quality and whether they were able also to detect the race of the speakers. While the exact details of this experiment were not provided, Morreira (2012:127) states that the results were inconclusive.

Morreira (2012:127) nevertheless affirms that this voice quality feature does indeed exist and expresses the wish that it be described in future and also avoids speculating about the origin of such a voice quality feature as a home language transfer phenomenon.

In a more recent perception experiment, Mesthrie, Chevalier and McLachlan (2015:392) tested two hypotheses, namely that crossing over towards the previously exclusively White SAE accent norms has been taking place amongst those Black students who have attended ex-model-C as well as private schools and secondly, that females are taking the lead in crossing over.

Mesthrie, Chevalier and McLachlan (2015:392) used 20 sound clips no longer than 12 seconds in duration in the perception experiment, taken from interviews of 13 middle-class Black (potentially crossover) speakers (8 females and 5 males), as well as a control of 7 speakers (2 male and 2 female White SAE speakers, 1 Black speaker selected as a distractor for gender and a further 2 speakers of traditional Black SAE, 1 male and 1 female). Mesthrie, Chevalier and McLachlan (2015:392) anticipated that the respondents would use both suprasegmental features (including stress, intonation, articulatory setting and rhythm) as well as segmental features in order to arrive at a holistic judgement.

Respondents were asked to judge from the sound clips (which were repeated twice) whether a given speaker was White or Black or otherwise to give an indication that they were unsure of the speaker's race (Mesthrie, Chevalier and McLachlan 2015:392). As distractor categories, respondents were also asked to judge speaker sex (whether male or female) as well as speaker age (namely, whether over the age of 30 years or below) (Mesthrie, Chevalier and

McLachlan 2015:393). The respondents for this perception experiment consisted of first-year Linguistics students, 151 in total, which after excluding certain respondents[22] consisted of 127 South African respondents (Mesthrie, Chevalier and McLachlan 2015:393).

Mesthrie, Chevalier and McLachlan (2015:404) found that listeners were better able to differentiate the ethnic identity of female speakers in comparison to male speakers. Among their findings, they also found that a larger percentage of Black listeners accurately judged the 7 crossover speakers to be Black in comparison to the percentage of accurate judgements for these speakers by White listeners (Mesthrie, Chevalier and McLachlan 2015:402). Mesthrie, Chevalier and McLachlan (2015:402) note that if the results could be generalized from this observed pattern, it may be the case that subtle articulatory setting and phonation effects are involved.

Mesthrie (2017) explicitly mentions articulatory settings as one of the suprasegmental features (along with rhythm and intonation patterns) which are employed in different ways by the speakers, but states that a direct analysis of phonation in particular is unwarranted at this stage, since SAE scholarship is not sufficiently advanced at present. The current research, it should be noted, is in part, a starting point in addressing this current state of affairs with regards to voice quality in SAE scholarship.

2.7. VOICE QUALITY IN ISIXHOSA

In comparison to the paucity of comment on voice quality variation in South African English in the literature, there is more extensive research investigating such phenomena for other South African languages. Given that all the black speakers included in my study were of an isiXhosa language background, and due to the possibility that some voice quality differences could plausibly be the result of L1-transfer effects, in this section I provide a brief discussion of the most relevant research relating to voice quality variation in isiXhosa.

[22] Some respondents were excluded for one of two possible reasons. Firstly, respondents would be excluded if they were not South African and secondly, if the respondents failed to identify the two traditional Black SAE speakers included in the recordings as Black (thus indicating either that they had not paid attention or that their intuitive grasp of sociolinguistic differentiation in their own society was unreliable).

IsiXhosa has been documented as containing a large number of breathy voiced consonants, as well as several ejective and aspirated consonants (see Finlayson 1989 for a full inventory).

Hundleby (1964) was one of the earlier researchers to document Xhosa English pronunciation. Hundleby (1964) describes varying levels of greater aspiration associated with certain consonants (for example, /p/ and /t/) in Xhosa-English pronunciation than in RP and also describes varying levels of devoicing for certain other consonants (for example, /g/ and /z/). The phoneme /h/, according to Hundleby (1964), is particularly in the English speech of 'less educated XEP' replaced with the substitute phone [ɦ], which could be described "as a vowel with glottal murmur."

More recently, Wissing (2002:141) has stated that if there are any perceptible differences between English speakers with different Bantu language backgrounds, that these would most likely be on the suprasegmental level.

Jessen and Roux's (2002) research is one of the few studies to use acoustic methods to investigate voice quality in isiXhosa, focusing particularly on contrasts associated with isiXhosa consonants. Jessen and Roux (2002:2) measured the correlates of breathy voice, in particular, as represented by the normalized spectral balance measure H1*−H2*, as well as an indicator of tonal depression, namely $f_{0.}$ Jessen & Roux (2002:9) recorded four Xhosa speakers reading specially designed wordlists and took measurements at five points within each vowel.

They found significant effects for Speaker, Category (defined as click accompaniment and stop type, namely plain, aspirated or voiced) and the interaction between Speaker and Category for all of the five measured periods for f_0 (Jessen and Roux 2002:15). They also found a significant result where f_0 was lower following voiced stops and clicks when compared to their plain or aspirated counterparts (Jessen and Roux 2002:17). According to Jessen and Roux (2002:17), these differences in f_0 are maintained essentially for the first part of the vowel up to the midpoint at least. Similar patterns were found for F1 (Jessen and Roux 2002:18).

The H1*−H2* parameter showed a similar pattern to that of f_0, thus with only a few exceptions for the last measurement point, speaker, category and interaction between them were found to be significant (Jessen and Roux 2002:21). Jessen and Roux (2002:22) observe that the speakers can be divided into two groups based on the patterning of the data for H1*−H2*. For one group, the values for H1*−H2* were lower following voiced stops and clicks when compared to the H1*−H2* values for the plain and aspirated stops and clicks. Thus Jessen and Roux (2002:22) assert that there is no indication that the voiced category of clicks and stops induces breathy voice on the following vowel for this group of speakers.

For the other speaker group, the H1*−H2* values are high, by comparison, following the voiced clicks and stops (Jessen and Roux 2002: 22). For this group, consisting of five speakers, Jessen and Roux (2002:22) claim that there is some indication of 'breathy voice.' They found that for the H1*−A3*[23] parameter, greater values predominated following both voiced and aspirated clicks and stops in comparison to following plain stops and clicks, but without as many significant results as was the case with the parameter H1*−H2* (Jessen and Roux 2002:30).

According to Jessen and Roux (2002:31), the results which they found for the parameter H1*−A3* agree with one of the main claims in the literature, namely that there is some degree of breathy voice following voiced clicks and stops. Evidence for this can be found in the observed levels of H1*−A3* following voiced clicks and stops, which are similar to the levels following the set of aspirated clicks and stops and which are also greater than the levels observed following the plain click and stop set (Jessen and Roux 2002:31).

Jessen and Roux (2002:36) point out that the acoustic property which varies the least from speaker to speaker, is the most statistically robust and most stable temporally when comparing plain and aspirated stops and clicks on the one hand and voiced stops and clicks on the other, was that of f_0 in the following vowel.

[23] The subtraction of the amplitude of the strongest harmonic in the third formant region from the amplitude of the first harmonic, both corrected for the effects of formants and their bandwidths on harmonic amplitudes. This measure as well as the other measures mentioned in this section will be discussed in more detail in the following section.

With only a small number of exceptions, f_0 values were found to be significantly reduced for voiced stops and clicks in the subsequent vowel, in comparison to the plain or aspirated cognates (Jessen and Roux 2002:36). Jessen and Roux (2002:36-37) however claim that this observed difference in isiXhosa has undergone phonologization to a greater extent than is the case in other languages.

Since 'breathy phonation' was found to vary considerably from speaker to speaker, Jessen and Roux (2002:38) suggest that in isiXhosa, it is not independently targeted, but rather that it is a concomitant effect of the gestures aimed at effecting f_0 depression, which is actively targeted. Jessen and Roux (2002:39) thus hypothesize that the primary voiced click and stop gesture in isiXhosa is that of larynx-lowering, which in turn leads to a slackening of the vocal folds, which is accompanied by 'glottal leakage' with potential breathy voice and strong lowering of the fundamental frequency and some lowering of F1. Thus the voice quality following the clicks and voiced stops of isiXhosa can best be characterized as 'slack voice' (Jessen and Roux 2002:40). Note that Jessen and Roux's (2002) study does not focus on indexical issues and is primarily focused on the phonological use of voice quality, unlike the current study.

2.8. REVIEW OF THE ACOUSTIC MEASURES USED IN THIS STUDY

In this section, I present a review of the various acoustic measures offered by VoiceSauce which are used in this study and offer a brief description of the use of the various measurements in prior research on voice quality variation and the perceptual correlates of these measures in terms of voice quality in so far as these have been established, thereby illustrating the relevance of the measures used for the acoustic investigation of voice quality differences. Relatively simple definitions of these measures as they are implemented in VoiceSauce are provided in section 3.5 (pages 98-102) in the following chapter. In this study, all of the measures described in this section are used because all may be relevant for the study of voice quality. This is because, as Keating and Esposito (2006) and Garellek (2016) demonstrate, different parameters have been found to be important for speakers of different languages. Not knowing which may be important for South African listeners necessitated the use of all available measures in the current study.

Harmonic Differential Measures

VS provides a number of harmonic differential measures. These measures derive their utility for the acoustic study of voice quality variation from observations that different voice qualities are expected to exhibit certain predictable effects on the harmonic structure of the speech signal.

2.8.1. H1−H2[24]

Fischer-Jørgensen (1967)

This measure is one of the most frequently used harmonic differential measures in studies of voice quality variation, particularly in sociophonetic studies as illustrated in the foregoing discussion. Fischer-Jørgensen (1967) was the first to recognize and demonstrate that the acoustic difference between breathy vowels and modal vowels in Gujarati could be accounted for in terms of their overtone component strength, using the harmonic differential measure H1−H2, that is, the difference between the amplitude of the fundamental component and that of the second harmonic (Esling and Edmonson 2011).

Bickley (1982)

Bickley (1982) found that this measure is successful in distinguishing breathy, modal and creaky phonation in !Xóõ and Gujarati.

Ladefoged and Antoñanzas-Barroso (1985)

Ladefoged and Antoñanzas-Barroso (1985) provide an explanation of why the measure H1−H2 may be useful for distinguishing breathy voice from modal voice, as observed in previous studies. The degree of vocal fold tension involved in the production of breathy voice is not as great as for modal voice and therefore the glottal pulse will have neither an abrupt closure nor an abrupt opening gesture (Ladefoged and Antoñanzas-Barroso 1985). This lack in a sharp airflow discontinuity Ladefoged and Antoñanzas-Barroso (1985) explain, will result in more energy for the fundamental as well as a reduction in energy for the upper harmonics.

[24] Note that for most of the harmonic measures provided by VS, corresponding corrected measures are also provided, which correct for the effects that formants and their bandwidths have on the harmonic amplitudes. Thus, for example, VS offers both H1−H2, as well as the corrected measure H1*−H2*, where by convention, the asterisks indicate that the respective harmonic amplitudes have been corrected for the influence of formants and formant bandwidths.

Huffman (1987)

Huffman (1987) found that this measure is successful in distinguishing breathy and modal vowels in Hmong.

Klatt and Klatt (1990)

Klatt and Klatt (1990) provide a more in-depth discussion of this measure. They point out that for modal vowels, there is simultaneous closure along the length of the vocal folds which adds an abrupt arrest to the airflow as well as a fairly strong higher harmonic excitation at the moment of vocal fold closure (Klatt and Klatt 1990). However, this is not the case for breathy vowels. In this case, the vocal folds first close anteriorly, with posteriorly propagating closure which adds a rounding of the corner of the glottal volume velocity waveform during closure (Klatt and Klatt 1990). Klatt and Klatt (1990) illustrate this corner rounding by means of a diagram, reproduced in figure 2.1 below.

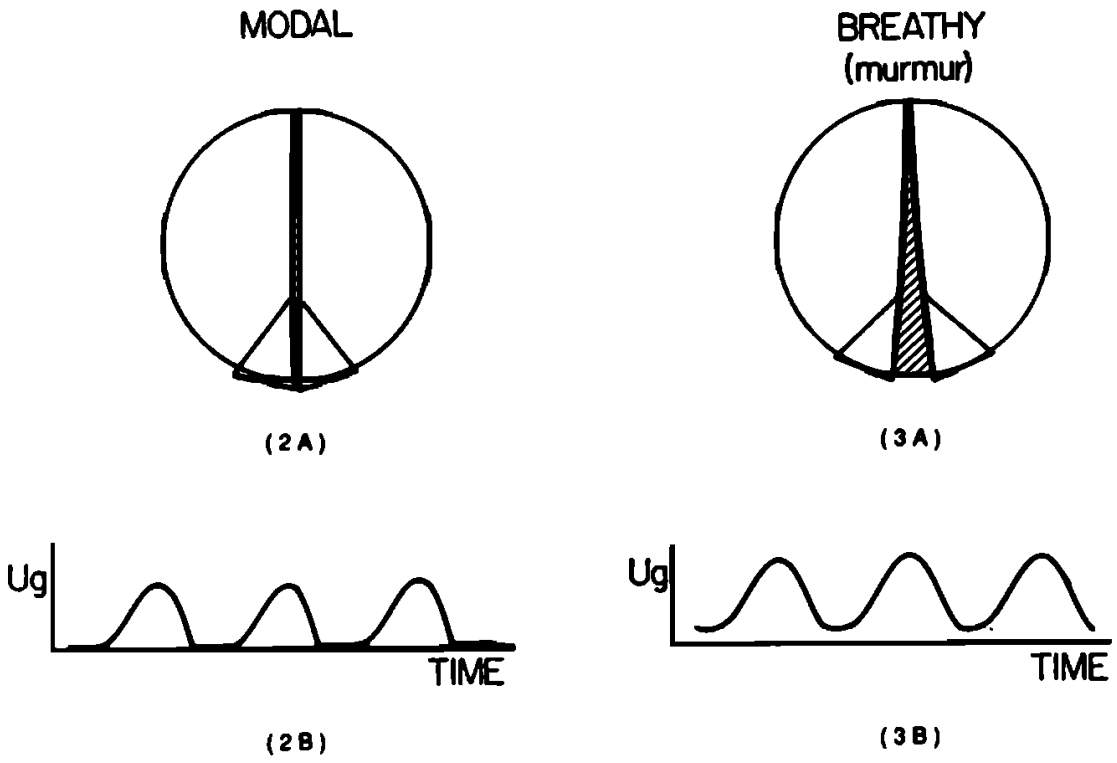

Figure 2.1: A comparison of the volume velocity waveforms for modal and breathy voice taken from Klatt and Klatt (1990:822) where Ug signifies glottal airflow.

As can be seen from figure 2.1 above, for modal voice (2B in figure 2.1) there is a relatively flat corner of the waveform during the closed phase of the glottal cycle. In contrast, the volume velocity waveform for breathy voice (3B in figure 2.1) displays a rounding of the corner during the closed phase.

According to Klatt and Klatt (1990), there are two main implications that this behaviour has on the harmonic components of the source spectrum. Firstly, the amplitude of the fundamental is increased due to the more sinusoidal shape of the waveform[25]. Secondly, because the closure is not simultaneous, there is substantial attenuation of the higher harmonic amplitudes (Klatt and Klatt 1990).

Based on their discussion of these effects, Klatt and Klatt (1990) suggest potential spectral cues which could be used to distinguish perceptually breathy vowels, such as the relative increase in the fundamental component amplitude as well as the higher harmonics being replaced by aspiration noise, both of which are captured by the measure H1−H2 (i.e a comparison of the relative strength of the fundamental component to that of a higher harmonic).

Holmberg, Hillman, Perkell, Guiod and Goldman (1995)

Holmberg et al.'s (1995) results suggest that in cases where inverse filtering[26] is unsuccessful and therefore measurements of adduction quotient based on the glottal waveform are unreliable, the harmonic differential measure H1−H2 could be used as a substitute for measuring open quotient[27] (OQ henceforth).

[25] Stevens (1998:68-69) however notes that for the lower frequency region of the spectrum, modal voice and breathy voice are approximately equal in terms of amplitude, but agrees with Klatt and Klatt (1990) that there will be attenuation of higher frequencies for breathy voice. This is because the higher frequencies of the spectrum are determined by the negative pulse of the derivative of the volume velocity waveform, which for a breathy vowel, rises less abruptly from the negative peak, thus leading to a greater attenuation of higher frequency amplitudes (Stevens 1998). For full details, see Stevens (1998).

[26] Inverse filtering is a technique whereby the process of speech production is effectively reversed (Ní Chasaide and Gobl 1997). As Ní Chasaide and Gobl (1997:429) note, this involves the passing of the speech signal through a filter, the transfer function of which is the "inverse of the supraglottal transfer function." In theory, this process cancels out the vocal tract's filtering effect (Ní Chasaide and Gobl 1997).

[27] Open quotient can be defined as "the fraction of time the glottis is open within one period" (Esling and Edmonson 2011:143).

Stevens and Hanson (1995)

Stevens and Hanson (1995) quantified the change in OQ, noting that when it varies from 30 percent to 70 percent, there is a change of approximately 10 dB in H1−H2. These authors therefore claim that the corrected measure H1*−H2*[28] provides a rough indication of OQ (Stevens and Hanson 1995).

Blankenship (1997)

Blankenship (1997) notes that since breathy vowels typically have a greater OQ (than say, modal or creaky vowels), one could predict that breathy vowels will therefore have larger values for H1−H2 than modal or creaky vowels would (Blankenship 1997). Blankenship (1997) also points out that it is often the case that for a given speaker and for vowels of a similar quality, smaller H1−H2 values will be found for laryngealised (i.e. creaky) vowels in comparison to modal vowels.

Hanson and Chuang (1999)

Hanson and Chuang (1999) concluded that female speakers have greater OQ's than male speakers. This conclusion was based on the observed greater values for H1−H2 for female speakers in comparison to male speakers on average (Hanson and Chuang 1999). Both female speakers and male speakers in these studies were native speakers of American English (with almost half of the female group having grown up in New England).

Keating and Esposito (2006)

Keating and Esposito (2006) provide a more detailed list of languages for which it has been demonstrated that harmonic differential measures (and H1−H2 in particular) can be used to distinguish between contrastive phonation types. There have been a few studies which have investigated the perception of differing values of H1−H2. These studies, according to Keating and Esposito (2006) have for the most part confirmed the perceptual importance of the harmonic

[28] As mentioned in the first chapter and explained with reference to the VoiceSauce measures discussed in the methodology section, it is expected that formants and their bandwidths would have an influence on harmonic amplitudes. For this reason, it is useful to correct the harmonic differential measures for the expected influence of formants and their bandwidths.

differential measure H1−H2 for the purposes of capturing cross-linguistic differences in the perception of phonation.

Keating and Esposito (2006) note that OQ could arguably be related to 'overall glottal stricture' and therefore, the measure H1−H2 could be suitable for characterizing any differences along the 'glottal constriction continuum,' as conceptualized by Ladefoged (1971) and Gordon and Ladefoged (2001).

Simpson (2009)

Simpson (2009) provides commentary of particular relevance to the discoveries regarding sex differences for this measure. As Simpson (2009) points out, citing Stevens, Fant and Hawkins (1987) and Maeda (1993), the first nasal formant occurs within the same region (200 Hz-350 Hz) for both female and male speakers. However, for a typical male f_0, the second and third harmonic components will be enhanced, whereas at a fundamental frequency typical for a female speaker, the nasal formant is more likely to affect the first harmonic (Simpson 2009). Since the H1−H2 measure is usually used for open vowels (in order to minimize the effects of the first formant and first formant bandwidth on the measure) and since the velopharyngeal port is more likely to be open during the production of open vowels, differences in terms of breathiness between male and female speakers measured solely in terms of H1−H2 may thus be confounded by the effect of the nasal formant, which is expected to be different for male and female speakers respectively (Simpson 2009). In my study, in addition to using the corrected equivalent of this measure, only female speakers are included as explained above, thereby allowing for a comparison between groups.

Esposito (2010)

Esposito (2010) found a correlation between perceptually-based judgements and H1−H2 and found that it was one of the commonly used ways of producing differences in phonation cross-linguistically. Esposito (2010) therefore also suggests that the reason underlying the effectiveness of this measure in capturing these phonation contrasts in so many languages may be due to the perceptual relevance of OQ for human listeners, suggesting that differences in OQ may be naturally salient.

Keating, Garellek, Khan and Kuang 2011

Keating, Garellek, Khan and Kuang (2011:1047) found that the measure H1*−H2* successfully distinguished phonation types in Gujarati, White Hmong, Southern Yi and Jalapa Mazatec such that the highest values for this measure were found for breathy/lax voice in each of these languages. They found that Contact Quotient (CQ) patterned together with H1*−H2* in distinguishing languages for which electroglottographic (EGG) data was available (Keating, Garellek, Khan and Kuang 2011:1047). Keating, Garellek, Khan and Kuang (2011:1049) conclude that this measure is the most important measure for distinguishing phonation types cross-linguistically.

Szakay (2012)

Szakay (2012) successfully used this measure to investigate ethnic differences in voice quality specifically, as discussed earlier in this chapter.

Chen, Park, Kreiman and Alwan (2014)

According to Chen, Park, Kreiman and Alwan (2014), the simplistic correlation between glottal OQ and the measure H1*−H2* is not consistent with evidence from more recent studies, which suggest that glottal gap (a gap that remains even during the closed phase of the glottal cycle), f_0 and glottal pulse skewness[29] may all have an influence on H1*−H2* values. However, while these authors demonstrate that the correlation is not as simple as previously thought, they do provide evidence that in the absence of a glottal gap, there is nevertheless quite a strong correlation between OQ and this measure and the correlation is reasonable overall.

Szakay and Torgersen (2015)

Szakay and Torgersen (2015), as discussed earlier in this chapter, used this measure to characterize regional and ethnic differences in voice quality in London English, concluding that non-Anglo speakers from inner-city London exhibit evidence of breathier phonation in contrast to Anglo speakers from Outer London.

[29] Skewness refers to "how rapidly the airflow closes off before the closed phase begins" (Esling and Edmonson 2011:143).

Keating, Garellek and Kreiman (2015)

Keating et al. (2015) point out that the measure H1−H2 (both corrected and uncorrected) is the most commonly used measure for creak and note that it provides a general reflection of glottal constriction, such that a lower value for this measure indicates a greater degree of constriction. They observe that in general, values for this measure will be low for most types of creak, since in most cases, creak involves glottal constriction (Keating et al. 2015). However, they also note that for one type of creak, namely 'non-constricted creak' (see Slifka 2000, 2006) the values for H1−H2 would be expected to be higher than those of modal voice.

2.8.2. H2–H4

Esposito (2006)

Esposito (2006) found that the H2−H4 measure was successful in distinguishing breathy from non-breathy phonation in several languages, including Chong, Fuzhou and Mon. Esposito (2006) lists Kreiman and Gerratt (2006) as having been one of the first to use this measure, where it was used for measuring vocal pathology.

Keating and Esposito (2006)

Keating and Esposito (2006) have used H2*−H4* in investigating phonation contrasts in certain languages, such as in Bura, for which they observed that the values for this measure are greater for low tones. This was interpreted to mean that (in conjunction with observed higher H1*−A3* values) such low toned vowels are breathier[30].

Kreiman, Gerratt and Antoñanzas-Barroso (2007)

Kreiman, Gerratt and Antoñanzas-Barroso (2007) investigated the usefulness of this measure in characterizing differences in spectral tilt in the higher frequency range by means of synthesis. They did not find a significant correlation between this measure and factors relating to spectral shape, unlike for H1−H2 (Kreiman, Gerratt and Antoñanzas-Barroso 2007).

[30] Higher values for these measures are a reflection of greater spectral tilt. As explained in the previous section dealing with H1−H2, spectral tilt is expected to increase with increasing breathiness.

Kreiman, Garellek and Esposito (2011)

These authors found that higher source H2−H4 values are used by White Hmong listeners in identifying phonemic breathy phonation using a synthesized vowel as a stimulus. The H2−H4 measure was found to significantly contribute towards the perception of contrastive breathiness independently of other measures, with increases in this measure being associated with a higher number of "breathy" as opposed to "modal" responses (Kreiman, Garellek and Esposito 2011). They also found that the effect for any increase in either H1−H2 or H2−H4 can be cancelled out when there is covariation between these two components, for example, when H1−H2 is increased simultaneously with H2−H4 being decreased (see further Kreiman, Garellek and Esposito 2011). Kreiman, Garellek and Esposito (2011) observed that for H2−H4, across different voices, the range of variability for H2−H4 values was approximately equal to 10.4 dB, with a just noticeable difference/range of approximately 0.29.

Keating, Esposito, Garellek, Khan and Kuang (2011)

These authors found that this measure did not distinguish the phonological phonation contrasts in Southern Yi, White Hmong, Jalapa Mazatec and Gujarati (Keating, Esposito, Garellek, Khan and Kuang (2011:1047). They suggest, based on their findings, that while it is unlikely that this measure contributes towards distinguishing between phonation types within languages, it is more likely that it differs across recordings, speakers and languages (Keating, Esposito, Garellek, Khan and Kuang 2011:1049).

Zhang, Kreiman and Gerratt (2011)

Using a vocal fold model, these authors found that there is an association between higher source[31] H2−H4 values and reduced body-layer stiffness of the vocal folds (Zhang et al. 2011).

[31] The term 'source' generally refers to the glottal source and 'source values' to the acoustic attributes of the glottal wave, traditionally obtained by means of inverse-filtering where the effect of the supraglottal vocal tract is removed. In Zhang's studies involving physical models of the vocal folds, source values are obtained by recording the acoustic signals produced by the physical model (under various conditions), normalizing these for amplitude, resynthesizing them, and then using an analysis-by-synthesis approach to assess the acoustic attributes, involving downsampling and then inverse filtering the downsampled signals using Javkin et al's (1987) method. For a detailed description of these methods, see Zhang, Kreiman, Gerrat and Garellek (2013).

Bishop and Keating 2012

Bishop and Keating (2012:1103) note that the measure H2*−H4* is infrequently used in studies of voice quality. They cite Kreiman et al.'s (2007) finding that the H2−H4 measure, in a comparison including nineteen different measures accounted for 8.3% of the variance in a principal components analysis and state that the measure must therefore play some kind role in capturing aspects of individual voice quality distinct from those captured by the more commonly used measures (Bishop and Keating 2012:1103). Using the recordings from Laver (1980), Bishop and Keating (2012:1103) cite exploratory work indicating that there is an association between lower values for H2*−H4* and some types of 'falsetto.'

Bishop and Keating (2012:1103) note that higher H2*−H4* would have an association with decreased stiffness of the vocal fold body-layer possibly in addition to breathy voice, while lower values for this measure are associated with possibly increased stiffness of the vocal fold body-layer as well as falsetto. Using logistic regression modelling of listener's responses to data extracted from sustained vowels, Bishop and Keating (2012:1108) observed that after f_0, H2*−H4* was the next most important parameter used for speaker sex identification due to the significant interaction observed between this measure and f_0. On its own, no significant main effect for the H2*−H4* measure was observed, but the measure showed a significant interaction effect with f_0 which was most prominent for those fundamental frequencies below the group mean fundamental frequency of 297 Hz (Bishop and Keating 2012:1109). Bishop and Keating (2012:1109) observed that there was a greater likelihood of listeners judging stimuli as 'male' for higher H2*−H4* values, particularly for lower fundamental frequencies. Bishop and Keating (2012:1111) suggest that in the fundamental frequency range of around 150 Hz, higher values for H2*−H4* might not be a reflection of breathiness but rather the absence of creaky voice which female speakers may typically produce at lower fundamental frequencies. Bishop and Keating (2012:1111) conclude that H2*−H4*, particularly in terms of its interaction with fundamental frequency, is one of the more relevant voice quality measures used in the identification of speaker sex. They do however point out that the role of H2*−H4* in speaker sex identification requires further research in order to clarify which aspects of the voice this measure reflects (Bishop and Keating 2012:1111).

Garellek, Keating, Esposito and Kreiman (2013)

Garellek, Keating, Esposito and Kreiman (2013) observed that H2−H4 is a significant predictor of breathiness in White Hmong, such that increases in this measure significantly increased the probability of listeners perceiving the stimulus as breathy. They also found that for speakers of this language, the measure H2−H4 may be in a trading relationship with H1−H2, such that if H2−H4 increases while H1−H2 decreases, the effect of both of these measures appears to be cancelled out (Garellek, Keating, Esposito and Kreiman 2013:1085).

Garellek, Samlan, Kreiman and Gerratt (2013)

Garellek, Samlan, Kreiman and Gerratt (2013) found that H2−H4 is a significant predictor of H1−H2 values, such that higher values for H2−H4 tend to lead to a decrease in H1−H2[32]. They also found that f_0, H4−2kHz (the amplitude of the fourth harmonic from which the amplitude of the strongest harmonic at 2000 Hz is subtracted) and H1−H2 are all significant predictors of H2−H4, such that increasing f_0 is associated with an increase in H2−H4, lower H1−H2 with an increase and particularly for females, higher H4−2kHz values with decreasing H2−H4 (Garellek, Samlan, Kreiman and Gerratt 2013). They suggest that listener sensitivity to the H2−H4 measure should vary as a function of both H4−2kHz and H1−H2 (Garellek, Samlan, Kreiman and Gerratt 2013).

Kreiman, Gerratt, Garellek, Samlan and Zhang (2014)

Kreiman, Gerrat, Garellek, Samlan and Zhang (2014) include H2−H4 as one of the important spectral parameters used in the psychoacoustic model (described in more detail in section 2.9).

Garellek and Seyfarth (2016)

These authors found that the H2*−H4* measure was useful as one of the measures of spectral tilt characterizing creaky vowels for which lower values of this measure were observed (Garellek

[32] For a thorough exploration and explanation of the interrelationships of the spectral tilt measures, the reader is encouraged to consult Garellek, Samlan, Kreiman and Gerratt (2013). The results obtained by these authors suggest listener sensitivity to a particular parameter would vary as a function of its relationship to adjacent parameters and note that future work will focus on whether listeners are indeed sensitive to such spectral configuration variations. As noted in a number of places throughout this thesis, perceptual sensitivity to a particular parameter is influenced at least to some extent by the listener's linguistic background and thus we would expect to find different parameters having greater importance than others in signalling particular voice quality contrasts for different linguistic groups.

and Seyfarth 2016). This difference in terms of lower H2*−H4* values was found to be greatest for phrasal creak, although the measure did also capture some of the differences between non-glottalized and glottalized tokens (Garellek and Seyfarth 2016).

Zhang (2016a)

In his original study, Zhang (2016a) did not use H2−H4 specifically, he did include a similar measure, namely H1−H4 (that is, the amplitude of the first harmonic minus the amplitude of the fourth harmonic), finding that in general, as medial surface thickness of the vocal folds increases, so H1−H4 tends to decrease, suggesting a similar relationship with vocal fold thickness for the measure H2−H4. The H1−H4 measure was also found to exhibit a small effect for increasing glottal angle[33], except in cases where the subglottal pressure were close to phonation onset and where the glottal angle at rest was large (Zhang 2016a:1501).

Garellek (2016)

Garellek (2016) notes that due to the fact that Zhang (2016a) observed lower spectral tilt for all frequency bands associated with increases in vocal fold thickness, the measure H2−H4 should also be lower when vocal fold thickness increases.

2.8.3. H1−A1[34]

Ladefoged (1983)

Ladefoged (1983) used this measure in quantifying breathiness for vowels. Ladefoged (1983) concludes that this is a useful, highly significant and reliable way of measuring voice quality differences between modal voice and 'murmur' (or breathy voice), since for all cases, based on data derived from speakers of Xóõ, where this contrast between murmur and modal voice is phonemic this measure was found to be larger for voiced sounds in comparison to the corresponding murmured sounds.

[33] In Zhang's (2016a:1495) three-dimensional vocal fold model, the glottal angle, α, is defined as the angle formed by "the medial surfaces of the two vocal folds" and which controls the resting opening of the glottis. See further Zhang (2016a:1495).

[34] As explained on the following pages, H1 is the amplitude of the first harmonic (i.e. for F0), while A1 is the strongest amplitude in the region of F1. See further Garellek (2016:17).

Kirk, Ladefoged and Ladefoged (1984)

Kirk, Ladefoged and Ladefoged (1984) note that as a result of the looser mode of vibration of the vocal folds involved in the production of breathy voice, there is frequently incomplete glottal adduction for the entire glottal vibration cycle. This results in an increase in transglottal airflow rate and this is in turn expected to produce turbulent airflow which results in a larger number of randomly occurring high frequency components (Kirk, Ladefoged and Ladefoged 1984).

In the production of creaky voice, the vocal folds are tenser by comparison and they close rapidly during the vibratory cycle (Kirk, Ladefoged and Ladefoged 1984). This results in a sharper excitation pulse which exhibits greater energy for the higher harmonics (Kirk, Ladefoged and Ladefoged 1984).

Thus, Kirk, Ladefoged and Ladefoged (1984) reason that, if one were to distinguish creaky from breathy voice by acoustic means, it would be ideal to use a measure which is able to distinguish whether a particular increase among the higher frequencies of the spectrum can be attributed to the presence of additional components which result from semi-random airflow turbulence (as in the case of breathy voice) or whether a relatively sharp glottal excitation (as typically found for creak) has produced an increase in higher harmonic intensity. This is illustrated in figure 2.2 taken from Klatt and Klatt (1990), which shows the replacement of weaker higher harmonics by aspiration noise as a result of semi-random airflow turbulence. Thus breathy voice does not result in an increase in higher harmonic intensity while creaky voice typically does.

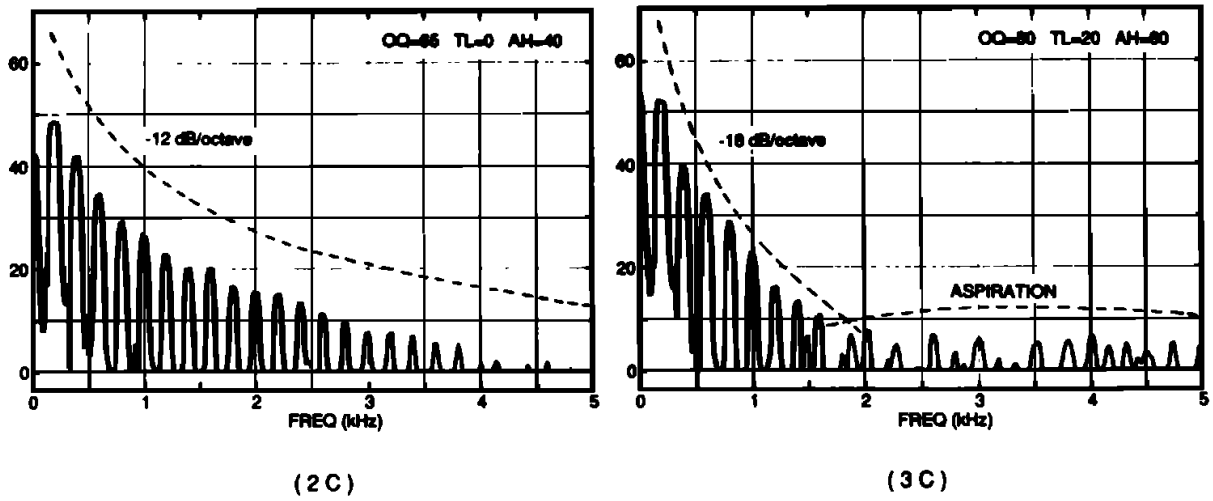

Figure 2.2: Comparison of spectra for a modal vowel (2C) and for a breathy vowel (3C) taken from Klatt and Klatt (1990:822).

Therefore, Kirk, Ladefoged and Ladefoged (1984) advocate the use of power spectra, since they enable the quantification of the relative energy found for different harmonics and chose the measure of the difference in decibels between the intensity of the fundamental component (i.e. H1[35]) and that of the most prominent harmonic in F1 (i.e. H1−A1).

Kirk, Ladefoged and Ladefoged (1984) found that breathy voice was successfully distinguished from modal voice using this measure for each speaker (of Jalapa Mazatec) in their study, since for any given speaker, the breathy voice value for this measure was greater than that for that speaker's modal or creaky voice (Kirk, Ladefoged and Ladefoged 1984).

Ladefoged and Antoñanzas-Barroso (1985)

Ladefoged and Antoñanzas-Barroso (1985) made use of an adapted measure of that used by Ladefoged (1983), namely, the difference in intensity between F1 and f_0 in decibels using a single Fast Fourier Transform (FFT) spectrum. The adapted measure used the FFT spectra produced at intervals of 10 milliseconds for the relevant vowel section (Ladefoged and Antoñanzas-Barroso 1985).They then calculated a mean for the spectra derived in this way and

[35] The first harmonic (or H1) is defined as "the lowest frequency sine wave component," that is, f_0 (Clark and Yallop 1990:203).

subsequently measured the difference between the fundamental and F1 in decibels for this average (Ladefoged and Antoñanzas-Barroso 1985).

Using the measure as described above, they found that for every speaker (of !Xóõ), there was a greater degree of spectral tilt in the lower spectral regions for breathy vowels when compared to modal vowels, where f_0 exhibited greater energy for breathy vowels by comparison (Ladefoged and Antoñanzas-Barroso 1985).

Ladefoged and Antoñanzas-Barroso (1985) claim that the differences in spectral tilt below 1000Hz are reflected to a great extent by this measure. They point out however that (uncorrected) H1−A1 would not be suitable when making a comparison between vowels which differ substantially in quality (Ladefoged and Antoñanzas-Barroso 1985).

Stevens and Hanson (1995)

Stevens and Hanson (1995) point out that 'glottal chink' size would have a negligible effect on spectral tilt, but would however clearly affect F1 bandwidth and therefore the measure H1*−A1[36] (Stevens and Hanson 1995). As they point out, F1 bandwidth relates to vocal tract energy losses (Stevens and Hanson 1995:151). This is because during the first part of every glottal period the sound-pressure waveform constitutes a damped oscillation for which the largest component is first formant frequency. F1 bandwidth can be shown to relate to the rate at which the amplitude of this damped oscillation decays[37].

Stevens and Hanson (1995:152) note that when the glottis is open such that there is transglottal airflow, glottal resistance contributes towards energy losses, thereby adding significantly to the bandwidth of the first formant. As A1 is the amplitude of the first formant, the measure H1*−A1 should therefore provide an indication of relative increases or decreases in F1 bandwidth.

Hanson (1997)

Hanson (1997) points out that one indication of the bandwidth of F1 is the amplitude of the peak of F1 in the spectrum. According to Hanson (1997), theory would predict that the amplitude of

[36] See footnote in section 2.8.1 on page 53 for the measure H1−H2 and for an explanation of the meaning of the asterisks for these measures.

[37] If the form of this oscillation is $e^{-\alpha t} \cos 2\pi f t$, where f signifies first formant frequency, then constant α (measured in sec^{-1}) relates to the bandwidth (BW) in terms of the following equation: $BW = \alpha/\pi$ Hz (Stevens and Hanson 1995: 151).

F1 in the speech spectrum should be proportionate to the bandwidth values when inversely transposed. Larger bandwidths therefore result in a reduction in the amplitude peak of F1, which therefore results in a decrease in the prominence of this peak in relation to the first harmonic amplitude (Hanson 1997). While formant bandwidths themselves may be difficult to measure, Hanson (1997) concludes that the measure H1−A1 may serve as a suitable indicator of F1 bandwidth. Hanson (1997) does however add that the method averages bandwidth over the whole glottal cycle, including the time during which the glottis is abducted. Both the bandwidth of F1 and F1 itself increase contemporaneously with the open phase of the glottal cycle and therefore using H1−A1 as a measure of F1 bandwidth may overestimate the bandwidth value during the first part of the cycle. Two other factors would impact on the measure, according to Hanson (1997:471), namely interspeaker variation with regards to first harmonic amplitude, as well as whether or not F1 happens to have its center 'on a harmonic'[38].

Based on her own data, Hanson (1997) found that the range of values for the measure H1*−A1 suggested that for some of the speakers sampled, F1 peaks were extremely prominent, but for others, these had undergone a high degree of damping. Hanson (1997) observes that the measure H1*−A1 correlates strongly with spectrally-based ratings of noise and exhibits a moderate correlation between estimated bandwidth for F1.

Hanson and Chuang (1999)

Hanson and Chuang (1999: 1066) agree with Stevens and Hanson (1995) that this measure can be used as an indicator of the presence of a posterior 'glottal chink'. According to Hanson and Chuang (1999), theory predicts that there will be a higher degree of correlation between the parameters H1*−A1, high frequency noise, as well as spectral tilt, if there is a posterior glottal chink present.

[38] The spacing of harmonics in the speech spectrum varies dependent on effective f_0 (Clark and Yallop 1990:205), with the spacing being wider for higher fundamental frequency ranges. Due to the continuous changes in both f_0 and resonance patterns in speech, there is no systemic and consistent relationship between harmonics and formant frequencies, such that these do not always coincide (Clark and Yallop 1990:225). When a formant is not centred on a harmonic, the amplitude of that formant is expected to be smaller than when centred on a harmonic (Hanson and Chuang 1999).

Based on the correlation between the measures H1*−A3* and H1*−A1, Hanson and Chuang (1999) concluded in agreement with studies such as that of Södersten and Lindestad (2009), that 'glottal chinks' are more common among females than they are among males.

Wayland and Jongman (2003)

Wayland and Jongman (2003) observed that in their data from Chanthaburi Khmer, there is a narrower bandwidth for F1 for clear vowels than is the case for breathy vowels, based on data for this measure. They claim therefore that H1*−A1 successfully distinguishes between clear and breathy vowels in Chanthaburi Khmer (Wayland and Jongman 2003).

Keating and Esposito (2006)

According to Keating and Esposito (2006), H1−A1 is one of the measures which can be used to distinguish between breathy and modal categories in a number of languages such as Mon and SADV (Santa Ana del Valle) Zapotec, even in cases where the measure H1−H2 (as described above) was not successful.

Esposito (2010)

Esposito (2010:308) found that H1−A1 is one of the measures which successfully distinguishes between breathy and modal phonation in Jalapa de Díaz Mazatec based on acoustic measures of Mazatec stimuli, where a discriminant analysis revealed that of seven measures, H1−A1 accounted for the second highest percentage of the variance (i.e. 20%). Esposito (2010:312) also found that Spanish listeners (as opposed to Gujarati and English listeners), weighed H1−A1 as a more important cue based on data from a similarity rating task than H1−H2 when considered individually. Esposito (2010:313) also used naturally produced stimuli from 10 languages with distinctions between breathy and modal phonation and based on acoustic measurements of vowel samples from these languages found that the measure H1−A1 was successful in distinguishing breathy from modal phonation in Fuzhou, Santa Ana Del Valle Zapotec and !Xóõ. Based on data derived from a free-sort task of these vowel samples, Spanish listeners showed a weak significant relationship between both H1−A1 and H1−H2 and listener judgements (Esposito 2010:314). For both the free sort task as well as for the similarity-rating task, Spanish listeners consistently weighed H1−A1 as a more important cue to perceptual breathiness than H1−H2.

Keating, Esposito, Garellek, Khan and Kuang (2011)

Keating et al. (2011:1047) found that this measure was successful in distinguishing between phonation types in Gujarati, Jalapa Mazatec and Southern Yi. However, using Multidimensional Scaling[39], they found that H1*−A1* does not relate strongly to any particular dimension distinguishing phonation types across these languages.

2.8.4. H1−A2

Bickley (1982)

Bickley (1982) found that ratings of breathiness by 6 native speakers of Gujarati who listened to synthesized vowels from Gujarati based on natural speech correlated with an increase in the H1−A2 values for these synthesized vowels.

Gobl and Ní Chasaide (1992)

Gobl and Ní Chasaide (1992) also document the behaviour of this measure (in the form of A2−H1) for lax, breathy, modal, whispery, creaky and tense voice. They observed that the amplitude of F2 was slightly weaker than that of the fundamental component for modal voice, a boosting of the amplitude of F2 relative to the fundamental for tense voice, an attenuation of F2 amplitude for lax voice in comparison to the amplitude of the first harmonic, with an even greater attenuation of F2 amplitude for both breathy and whispery voice (Gobl and Ní Chasaide 1992). Creaky voice was found to display fluctuating F2 amplitude levels (Gobl and Ní Chasaide 1992).

Blankenship (1997)

Blankenship (1997), based on the results of a pilot test, concluded that among the best measures for distinguishing between modal and laryngealised samples was the measure H1−F2 (i.e. H1−A2), but that this measure also worked well for distinguishing breathy from modal vowels.

[39] Multidimensional Scaling (or MDS) is an analysis technique in which the measured distances between items of interest are used to define a map where the spatial arrangements of those distances are represented. For a more detailed description, see further Keating et al. (2011:1048).

Consequently, Blankenship (1997) selected this measure for her main experiment, as a measure of the abruptness of vocal fold closure.

Blankenship (1997) explains why theoretically, this should be a measure of the abruptness of vocal fold closure. When there is gradual closure of the vocal folds over their entire length, there is an excitation of mainly the lower vocal tract frequencies, resulting in a sound wave which is closer to a sinusoid and dominated by f_0. The corresponding spectrum of a wave of this description will slope steeply downwards, such that there will not be much energy for the higher frequencies, but there will be far more energy in the region of f_0 (Blankenship 1997). However, when there is an almost instantaneous, rather than gradual closing of the vocal folds, a larger frequency range is excited, resulting in a more complex wave with energy at many more frequencies (Blankenship 1997).

The spectral slope has a gradient which is less steep for such a wave with a greater spread of energy for all of the frequencies (Blankenship 1997). In cases where there is no simultaneous closure of the vocal folds along their whole length, such as for example, in the case of breathy voice, it could be predicted that there would be high values for the measure H1−A2, while for phonation types where there is an abrupt adduction of the vocal folds as a result of vocal fold tension, such as in the case of laryngealisation, H1−A2 values would be expected to be smaller and perhaps negative rather than positive (Blankenship 1997).

Esposito (2006)

Esposito (2006) compared the acoustic measures for breathy and modal vowels in several different languages. A number of languages including Chong, Fuzhou, Mon, Santa Ana del Valle Zapotec, San Lucas Quiaviní Zapotec, Tlacolula Zapotec and Tamang were found to distinguish modal from breathy vowels based on differences in H1−A2 (Esposito 2006). Overall (for the ten languages studied), H1−A2 was found to be the fourth most successful of the measures for distinguishing breathy from modal phonation (Esposito 2006:45). Esposito (2006:91) also found that for three Mazatec speakers, a discriminant analysis revealed that H1−A2 accounted for most of the variance in the data between breathy and modal vowels out of the eight acoustic measures used. In a perception experiment using similarity-rating tasks based on stimuli from the three Mazatec speakers, Esposito (2006:140) found that none of the Gujarati, English and Spanish

listeners relied on the H1−A2 measure in making auditory judgements. This is spite of the importance of this measure as suggested by the results of the acoustic analysis where it accounted for a large portion of the variance in the data (Esposito 2006).

Keating and Esposito (2006)

Keating and Esposito (2006) note that H1−A2 is one of the measures which distinguishes breathy from modal categories in certain languages such as Mon and Santa Ana del Valle Zapotec, where the measure H1−H2 fails to distinguish these categories.

Keating et al. (2011)

Keating et al. (2001:1047) found that this measure was successful in distinguishing contrastive phonation types from one another in Southern Yi, Jalapa Mazatec and Gujarati, but was however found not to be related to any of the important Multidimensional Scaling dimensions they identified.

2.8.5. H1−A3

Holmberg et al. (1995)

As Holmberg et al. (1995) note, citing Stevens (1977), when airflow is abruptly reduced, there is an excitation of energy among the higher frequencies of the speech spectrum. In order therefore to find information in the speech spectrum which reflects such abrupt glottal waveform changes, Holmberg et al. (1995: 1213) used the measure of the difference in the fundamental amplitude and 'the highest spectral peak in the F3 region' (i.e. H1−A3).

Holmberg et al. (1995) formed various hypotheses with regards to the effects of different types of glottal action on the measure H1−A3. For gradual adduction, a more sinusoidal waveform is predicted and a decrease in adduction quotient and thus the spectrum would exhibit a comparatively high first harmonic amplitude, an increased spectral slope overall, as well as a lower sound pressure level and an attenuation of the peak of F1 (Holmberg et al. 1995). Therefore, these researchers predicted that for such closing movements of the glottis, there would be a negative correlation between H1−A3 and adduction quotient and between H1−A3 and sound pressure level (Holmberg et al. 1995).

Stevens and Hanson (1995)

Stevens and Hanson (1995) include H1−A3 as a measure of spectral tilt in their study. These researchers observe the large range in the spectral tilt of their subjects for this measure and link it to either non-simultaneous adduction along the entire length of the vocal folds, or a 'glottal chink' being present or the contribution of both of these factors (Stevens and Hanson 1995).

Hanson (1997)

Hanson (1997) claims that H1−A3 values may be taken to indicate spectral tilt of the source with reasonable accuracy, apart from cases where H1 is weak. Hanson (1997) also notes that there are other factors influencing third formant amplitude, for example first and second formant locations, as well as the radiation characteristic of the mouth which affects the third formant bandwidth more so than it does lower formant bandwidths, which is an effect which will be different for different vowel (and specifically lip) configurations and therefore notes the importance of using compensatory corrections to normalize for these effects for cross-vowel or cross-speaker comparisons when using this measure. If F3 is found to be 'centered on a harmonic', this will also affect the value of A3 and therefore the measure H1−A3 (Hanson 1997: 469).

Blankenship (1997)

Blankenship (1997:9) notes if there is a successful correlation between this measure and spectral slope, then it can be predicted that laryngealised vowels would provide small values for this measure, modal vowels would yield values of medium magnitude, while large values would be found for breathy vowels.

Hanson and Chuang (1999)

Hanson and Chuang (1999) take this measure to be a reflection of the spectral tilt of the source. This is because the greatest influence of the abruptness of airflow cut-off during glottal adduction is manifested in the mid-frequency to high-frequency range (Hanson and Chuang 1999). In their study, Hanson and Chuang (1999) corrected H1 for the first formant effect and corrected A3 for the effect of F1 and F2 on the prominence of the F3 amplitude peak. This is

taken as an important normalization procedure when comparing values for this measure across vowels of different qualities. This therefore yielded the corrected measure H1*−A3*. Hanson and Chuang (1999) found a high correlation between H1*−A3* as an indicator of spectral tilt and their measure evaluating noise[40] and their measure of the prominence of the first formant, namely H1*−A1. Hanson and Chuang (1999) also found that this was the only measure out of those which they included in their study to show a significant difference between different vowels.

Esling and Edmonson (2011)

Esling and Edmonson (2011) take the measure H1−A3 (divided by three octaves) to be a representation of the skewness of the glottal pulse, in other words, the rapidity of the closing off of airflow prior to the initiation of the closed phase of the glottal cycle.

Keating et al. (2011)

Keating et al. (2011:1047) found that this measure was successful in distinguishing contrastive phonation types for Gujarati, Jalapa Mazatec and Southern Yi. Using Multidimensional Scaling, Keating et al. (2011:1048) found that this measure (along with CPP and H2*−H4* and energy) relates to a dimension which both separates the four languages they examined as well as distinguishes between breathy/lax phonation and other phonation types based on noise and spectral tilt.

2.8.6. H4−H2K

Kreiman, Garellek and Esposito (2011)

Kreiman et al. (2011) found that the variability range for this measure across different voices was approximately 18.3 dB, while the just noticeable difference range was approximately 0.16 dB. The findings provided further confirmation of the perceptual importance of this measure (Kreiman et al. 2011). Kreiman et al. (2011) found that this measure is a significant contributor

[40] Hanson and Chuang (1999) used a measure of noise (N_w) based on Klatt and Klatt's (1990) measure which employs visual inspection of waveforms in order to estimate the noise component in the signal (thus achieving what a measure like HNR achieves by means of a formula by means of visual inspection instead). A four-point scale was used to rate vocalic waveforms, from 1 (signifying a waveform which was periodic and without visible noise) through to 4 (where noise is predominant and where periodicity is minimal or absent). See further Hanson and Chuang (1999) for details. For a description of the original noise rating measure, see Klatt and Klatt (1990).

to perceptions of contrastive breathiness, at least for listeners who were speakers of White Hmong.

Kreiman et al. (2014)

Based on the speech synthesis experiments of Krieman et al. (2011) and Kreiman and Gerratt (2011), Kreiman et al. (2014) observed that greater detail would be needed for modelling the spectrum of the source above H4. Therefore, they added the measure H4−2K (in addition to 2K−5K, discussed below) and found that listeners were not able to distinguish between synthetic and natural tokens consistently, concluding that this parameter should be included in the psychoacoustic model (described in the following section).

Garellek, Samlan, Gerratt and Kreiman (2016)

Garellek et al. (2016:1408) found that neither the 2K−5K slope nor that of H2−H4 significantly changed the just noticeable differences for the H4−2K measure. They point out that when the slope of H4−2K is comparatively flat, either noise level changes or overall spectral slope changes will influence perceptions of voice quality, while if the H4−2K slope is steeper, changes to those frequencies above 2 kHz are not as important in terms of perception (Garellek et al. 2016:1409). H4−2K is mentioned as one of the robust components of the psychoacoustic model (Garellek et al. 2016:1408).

Garellek (2016)

Garellek (2016) notes that H4−2K values should decrease as vocal fold thickness increases.

2.8.7. 2K–5K

Kreiman, Garellek and Esposito (2011)

Kreiman, Garellek and Esposito (2011) found this measure to be a significant contributor towards the perception of phonemic breathiness at least for White Hmong-speaking listeners.

Zhang, Kreiman, Gerratt and Garellek (2013)

Using a physical model of the vocal folds, Zhang et al. (2013:460) found that in asymmetric conditions of the vocal folds in terms of stiffness, the 2K−5K measure correlates with left vocal fold stiffness, such that with increased stiffness, this measure increases.

Garellek et al. (2016)

Garrellek et al. (2016:1408) found that 2K−5K is the only parameter of the spectral source model which is not independent perceptually from the other parameters of the model, since H4−2K significantly interacts with 2K−5K, with steeper H4−2K being associated with a higher just noticeable difference in 2K−5K. They also found that the presence of noise also significantly affects the just noticeable difference in 2K−5K values (Garrellek et al. 2016:1408). Garellek et al. (2016:1409) discovered that the just noticeable difference for 2K−5K was only affected by noise when H4−2K was flatter and that the slope of this latter component only had an impact on the 2K−5K just noticeable difference when noise was lower. They thus conclude that the sensitivity of listeners to the 2K−5K parameter is dependent on an interaction between the spectral slope (particularly for H4−2K) and levels of noise (Garellek et al. 2016:1409).

Garellek (2016) points out that the extent to which linguistically-relevant distinctions in voicing are signaled by 2K−5K (as well as H4−2K) is not yet established. Garellek (2016: 28) also predicts, based on Zhang (2016a) that 2K−5K should lower as vocal fold thickness increases.

2.8.8. Measures of Signal Aperiodicity in VS

VS provides Harmonics-to-Noise ratio (HNR) and Cepstral Peak Prominence (CPP) measures which both capture two different types of noise which may be present in the source, namely aspiration noise as well as noise which arises when pitch is poorly defined. As pointed out by Garellek (2016), while such measures do not define the presence or absence of breathy voice on their own, when used in combination with spectral measures (such as H1−H2) which provide an indication of whether the glottis is relatively constricted or relatively spread, such measures can be useful in describing voice quality.

2.8.8.1. HNR (Harmonics-to-Noise Ratio)

Yumoto, Gould and Baer (1982)

Yumoto, Gould and Baer (1982:1544) developed an H/N (harmonics-to-noise) ratio formula and assessed its usefulness in the quantitative assessment of pathological hoarseness. They found strong correlations between subjective hoarseness severity ratings and this measure and also found that there are substantial differences between non-pathological and pathological research subjects in terms of this measure. They conclude that the formula was successful at assessing the phonatory noise component and that it is a useful measure for quantitatively assessing a vowel's noise component relative to the vowel's harmonic component (Yumoto, Gould and Baer 1982:1549).

Hillenbrand (1987)

Hillenbrand (1987) tested Yumoto et al.'s (1982,1984) HNR technique in terms of accuracy by means of simulation and using automatic measurements of HNR by means of a program. Hillenbrand (1987:452) found that this measure is unable to differentiate between additive noise and perturbation[41]. Hillenbrand (1987:453) also found that at favourable HNR values, there were few pitch-tracking errors and they were relatively small in size when they did occur, but that the algorithm was particularly sensitive to these errors. However, with decreasing HNR, there is less sensitivity to measurement errors, but the size as well as the number of those errors increased (Hillenbrand 1987:453). Hillenbrand (1987:454) concluded that HNR is strongly affected by pitch perturbation variations. Changes in values for HNR are therefore difficult to interpret, since they may be attributed to either increases in amplitude perturbation, pitch perturbation, additive noise or a combination of all of these (Hillenbrand 1987:455;456). Hillenbrand (1987: 455) found that jitter may independently influence HNR measurements and also that shimmer tends to decrease HNR. Hillenbrand (1987:457;458) also suggests a couple of modifications to Yumoto et al.'s (1982) HNR technique, namely normalizing for differences in the amplitude of the pitch

[41] According to Hillenbrand (1987:448), additive noise generally refers "to the acoustic by-product of turbulent air flow generated at the glottis during phonation." Perturbation on the other hand, covers both jitter (cycle-to-cycle fundamental frequency variations) and shimmer (cycle-to-cycle amplitude variations) (Hillenbrand 1987).

pulse for RMS intensity[42], as well as setting the averaging window size according to the smallest period.

de Krom (1993)

De Krom (1993) set out to formulate a novel harmonics-to-noise ratio calculation technique by employing the harmonic frequency distribution in 'the log magnitude spectrum' in defining the noise. De Krom (1993) defines a comb-filter in the cepstral domain in order to separate the noise from the harmonics and thereby to calculate the HNR.

De Krom (1993) notes that in the calculation of this algorithm, the majority of the variance in his observed data for HNR could be accounted for in terms of additive noise. Other variables of importance also influencing the HNR values include window length as well as fundamental frequency (de Krom 1993).

De Krom (1993) also observes that the influence of jitter on HNR levels is particularly strong even at very small levels of jitter, such that for a given f_0, there is an inverse relationship between the percentage of jitter and HNR.

De Krom (1993) defines the noise spectrum as the base on which the harmonics of the spectrum lie. Having obtained the noise spectrum thus defined, it is then possible to calculate the HNR by differencing the level between the noise spectrum and the original spectrum for any chosen frequency band (de Krom 1993). The novel part of this operation which de Krom (1993) introduced is the 'rahmonic comb-liftering operation' which is performed in the real cepstral domain in order to derive the noise spectrum[43].

De Krom (1993) found by means of experiment that there is a strong negative correlation between both the noise burst level and HNR and the extent of excitation signal frequency perturbation, that is between HNR and jitter as well as added noise, with a smaller effect for f_0 and window length.

[42] In their study, Hillenbrand (1987) wrote a program which measured individual pitch pulse intensity and subsequently scaled all of the pitch pulses to the same RMS intensity value which they reasoned resulted in a substantial reduction in the effects which amplitude perturbation has on HNR values.

[43] Terms such as 'rahmonic,' 'liftering' and 'cepstral' are all technical terms used to indicate aspects of cepstral analysis (by analogy with 'cepstrum' which is the word 'spectrum' with the first syllable read back to front) (Kent and Read 2002:95). Thus for example, 'liftering' refers to a filtering operation in the cepstral domain.

Wayland, Gargash and Jongman (1994)

Wayland, Gargash and Jongman (1994) tested de Krom's (1993) HNR algorithm on the speech of three Javanese speakers derived from the reading of a word list consisting of breathy and clear pairs of words. They found that the algorithm was reliably able to distinguish the clear from the breathy tokens for all of their speakers, such that clear tokens had higher HNR values (Wayland, Gargash and Jongman 1994).

Walton and Orlikoff (1994)

Walton and Orlikoff (1994) used this measure in their investigation of speaker race identification in America and found that, in comparison to the white speakers in their sample, the black speakers exhibited a mean HNR which was significantly lower. According to Walton and Orlikoff (1994), this measure is a reflection of variability in amplitude and frequency cycle-to-cycle as well as reflecting 'additive noise.'

Wayland and Jongman (2003)

According to Wayland and Jongman (2003), it can be expected that the HNR values of breathy vowels should be comparatively low relative to those of clear vowels, particularly for F3 and formant frequencies above F3.

Wayland and Jongman (2003) found that HNR (using de Krom's 1993 algorithm) was not particularly useful in distinguishing breathy from clear vowels in Chanthaburi Khmer.

Esling and Edmonson (2011: 142)

Esling and Edmonson (2011) predict that there would be smaller values for HNR for creaky voice than for harsh voice, due to the greater modulation noise present for creaky voice.

Esling and Edmonson (2011: 142) link elevated HNR levels to the presence of modulation noise, the type of noise resulting from 'jet turbulence' (due to rapid transglottal airflow), collisions of the vocal folds, elevated levels of air pressure below the glottis, flaccidity of the vocal folds, over-adduction of the vocal folds, or from pathological causes, all of which result in 'non-linear motion' present during laryngeal vibratory activity.

Garellek and Keating (2015)

According to Garellek and Keating (2015), noise in the lowest of frequency ranges (reflected in the VS measure HNR05), represents either added noise or voicing irregularity. They found that for utterance final creaky vowels, HNR was lower, which they interpreted to mean that there is greater aperiodicity for utterance final creak (Garellek and Keating 2015).

Keating et al. (2015)

Keating et al. (2015) state that low values for HNR are an indication that the periodic excitation is weaker in comparison to glottal noise, which can result from either harmonics which are ill-defined (such as which result from irregular fundamental frequency) or from the prominence of noise generated at the glottis. They also point out however that for one type of creak, namely vocal fry, HNR is expected to display relatively high values, due to the sharp definition of glottal excitations involved in the production of this type of creak (Keating et al. 2015). Keating et al. (2015) point out that in English (citing studies by Garellek 2012, 2014 and 2015), Ju|'hoansi (Miller 2007), Hmong (Andruski 2006 and Garellek 2012), Taiwanese (Pan, Chen and Lyu 2011) and Mazatec (Garellek and Keating 2011), irregular fundamental frequency, as evidenced by low values for HNR is correlated with creaky voice.

2.8.8.2. Cepstral Peak Prominence (CPP)

The CPP measure is fairly recent in its use as a measure of voice quality.

Klatt and Klatt (1990)

These authors used subjective ratings by visual inspection to represent periodicity, which could be considered as the precursor to the automated CPP measure.

de Krom (1993)

The basic principle on which CPP is based is similar to that of de Krom's (1993) 'signal-to-noise ratio.'

Hillenbrand, Cleveland and Erickson (1994)

The first to introduce the measure of CPP in the assessment of different voice qualities was Hillenbrand, Cleveland and Erickson (1994) in their investigation of the acoustic correlates of breathy voice quality.

According to Hillenbrand et al. (1994: 772), CPP is a representation of the distance between the cepstral peak against 'background noise.' Hillenbrand et al. (1994) concluded that a substantial proportion of variance in breathiness ratings was accounted for by cepstrum-based measures of periodicity which they used, with CPP acting as an accurate predictor for both filtered and unfiltered speech signals.

Hillenbrand et al. (1994) claim that their CPP measurement appears not to be dependent on accurate f_0 tracking (unlike the HNR measure described above).

Hillenbrand and Houde (1996: 314)

According to Hillenbrand and Houde (1996: 314), a cepstrum can be defined as 'a log power spectrum of a log power spectrum'. Therefore, for signals which are periodic, the first log power spectrum will exhibit energy which will be spaced at frequencies which are harmonically related and the second log power spectrum of this spectrum will have a component which will be stronger when there is a greater regularity of the harmonic peaks in the first spectrum (Hillenbrand and Houde 1996). The cepstral peak, which is more prominent for signals with a well-defined harmonic structure occurs at a time (called the 'quefrency') which corresponds to the signal's fundamental period (Hillenbrand and Houde 1996: 314). CPP measures the amplitude of this cepstral peak which corresponds to the fundamental period, but is normalized for the overall amplitude of the signal (Hillenbrand and Houde 1996). The amplitude of this peak, according to Hillenbrand and Houde (1996) simultaneously represents the overall signal amplitude as well as the degree of the harmonic organization. In order to normalize for the overall amplitude of the signal, a linear regression line is used in order to relate the magnitude of the cepstrum to quefrency (Hillenbrand and Houde 1996). Therefore the measure of CPP is the difference between the cepstral peak amplitude and the value which corresponds to the regression line directly under the cepstral peak (Hillenbrand and Houde 1996). Since CPP reflects the extent to which the signal displays a harmonic structure, Hillenbrand and Houde

(1996) predicted that smaller CPP values would be found for breathier signals in comparison to non-breathy ones.

Esposito (2006)

Esposito (2006:40) found that CPP was successful in distinguishing breathy from modal vowels in 10 languages, including Chong, Fuzhou, Green Hmong, White Hmong, Mon, Santa Ana del Valle Zapotec, San Lucas Quiviní Zapotec, Tlacolula Zapotec, Tamang and !Xóõ. Esposito (2006:44-45) concluded from this that all of these languages used differences in noise in forming the distinction between phonemically breathy and modal vowels. Esposito (2006:52) also found that CPP accounted for most of the variance of the eight measures tested using a discriminant analysis. However, based on a listening experiment, Esposito (2006:76) found that Spanish listeners did not use CPP in a free-sort task, but instead used H1−A1 and H1−H2. This is contrary to the predictions based on the importance of CPP as suggested by the discriminant analysis (Esposito 2006:76). Likewise, there was no significant relationship between CPP and the auditory judgements of English and Gujarati listeners (Esposito 2006:77). Esposito (2006:77) suggests that this may possibly be attributed to the fact that the range for the stimuli in terms of CPP is relatively small for the data used in the discriminant analysis and that listeners may perhaps only be sensitive to larger differences in CPP. Esposito (2006:91) also found that CPP was a successful measure of phonation for distinguishing between breathy and modal vowels for three Mazatec speakers. This in turn would suggest that these speakers use noise in order to form these contrasts (Esposito 2006:91). In a similarity rating task however, in spite of not using CPP in the free-sort task, English listeners did use CPP to distinguish between breathy and modal stimuli (Esposito 2006:136-137). English listeners were found to rely primarily on H1−H2, followed by CPP in their similarity rating judgements (Esposito 2006:137). Esposito (2006:162) found evidence that (American) English listeners judge pathological stimuli according to both H1−H2 as well as CPP. While Klatt and Klatt (1990) as well as Hillenbrand et al. (1994) found that listeners relied most on aspiration noise (as reflected in CPP values in the case of Hillenbrand et al. 1994 and a subjective visual rating equivalent in the case of Klatt and Klatt 1990) in order to distinguish signals differing in breathiness, Esposito (2006:174) found on the contrary, that the listeners in her study relied on both H1−H2 and CPP to distinguish breathy stimuli, with H1−H2 being the more important of the two measures.

Keating and Esposito (2006)

As reported by Keating and Esposito (2006), CPP has been useful in distinguishing breathy from modal voice in a number of languages where this contrast is phonemic.

Keating et al. (2011)

These authors found that CPP distinguishes between contrastive phonation types in White Hmong, Jalapa Mazatec and Southern Yi, but not in Gujarati (Keating et al. 2011:1047). CPP is also related to a dimension which contributes to the distinction between breathy/lax phonation types from other phonation types based on noise and spectral tilt as well as contributing towards separating languages (Keating et al. 2011:1048).

Fraile and Godino-Llorente (2014)

Fraile and Godino-Llorente (2014) state that the assumption of a relation between the perception of breathiness and the measure CPP may be presently considered to be well-founded, but point out that the exact degree of such a correlation may be language-specific with regards to the speaker and may also be dependent on the listener's linguistic background. Fraile and Godino-Llorente (2014) also state that there is a relationship between CPP and the physiological processes which govern breathiness production although they note that CPP variations may result from a number of different vibratory patterns of the vocal folds, as well as other articulatory characteristics. This means that it is not feasible at present to identify and directly link any particular physiological cause for any specific change in CPP values (Fraile and Godino-Llorente 2014).

Fraile and Godino-Llorente (2014) conclude that measures of a number of features which describe aperiodicity in the signal are integrated by CPP. According to these authors, this is why there is a relationship between CPP and general dysphonia, but also why specific aspects of voice quality are not adequately predicted by the measure CPP alone.

Keating et al. (2015)

Keating et al. (2015) observe that through re-synthesis, it appears that CPP is lowered by adding jitter, due to an increase in noise and thus for types of creak with irregular f_0, lower CPP values are predicted.

2.8.8.3. Subharmonic-to-Harmonic Ratio (SHR)

Sun (2002)

Sun (2002) developed a perception oriented pitch-determination algorithm to be able to analyze speech signals for both normal speech as well as speech characterized by alternate pulse cycles. In discussing the theory behind the use of the measure, Sun (2002) notes that modulation is manifested in the frequency domain as subharmonics and points out that there is a close relation between pitch perception and SHR (that is the ratio in amplitude between subharmonics and harmonics).

Sun (2002) points out that above a certain SHR threshold, subharmonics appear clearly in the speech spectrum and there is a perceptual decrease in pitch by approximately one octave. This would suggest that an optimal way of determining pitch would be by means of an SHR computation (Sun 2002). Sun (2002) evaluated the performance of the newly proposed algorithm by using data from 50 sentences by a male and female speaker respectively and concluded that SHR is one of the more reliable methods for determining pitch and is superior to all other frequency domain-based methods. Sun (2002) also found, using synthesized speech materials, that the SHR algorithm performs better than another frequency domain-based measure and SHS[44], which was found to have a greater level of sensitivity to alternate pulse cycles and concluded therefore that the new SHR algorithm is useful in reducing the overall error rate for speech signals characterized by alternate pulse cycles (and therefore increased subharmonics).

Shue et al. (2011)

According to Shue et al. (2011), this measure may be particularly suitable for the characterization of speech marked by alternating pulse cycles.

[44] Subharmonic Summation algorithm. See further Hermes (1988).

Keating et al. (2015)

Keating et al. (2015) point out the usefulness of this measure in identifying certain sub-types of creak. They note that while most types of creak will not have high SHR values, higher values would be observed for multiply-pulsed creak (Keating et al. 2015). Keating et al. (2015) note that because multiply-pulsed creak has more subharmonics, it will be characterized by higher values for SHR.

Khan et al. (2016)

Khan et al. (2016) found that when comparing production data for various types of creak as produced by five transgendered male American English speakers, lower values for SHR were observed. In combination with higher values for most harmonic differential measures and lower values for HNR, Khan et al. (2016) were able to conclude that the type of creak being used by their research subjects could be characterized as non-constricted creak (discussed in more detail where relevant in chapters 5 and 6).

2.8.9. Energy (Root Mean Square Energy)

The Root Mean Square value is an expression of sound pressure which is proportional to the acoustic property of intensity, the way in which power (the rate of expended energy in producing sound) is distributed in space (Clark and Yallop 1990).

In calculating the RMS amplitude, the value for every sample occurring in the sampling window is typically squared, thereby exaggerating the differences as well as eliminating negative values (Kent and Read 2002). Thereafter, the average of these squared values is calculated, followed by calculating the square root of this mean in order to bring this value back to that of the original point of comparison (Kent and Read 2002).

Wayland and Jongman (2003)

Wayland and Jongman (2003) used a measure of vowel RMS amplitude in their investigation of the differences between clear and breathy vowels in Chanthaburi Khmer and discovered that for all except one speaker, RMS amplitude was significantly greater for breathy vowels in comparison to clear vowels, concluding that this measure was useful for signaling the difference between these two categories in the language.

Keating et al. (2011)

These authors found that energy distinguishes between phonation types in Jalapa Mazatec, but not in Southern Yi, Gujarati and White Hmong (Keating et al. 2011:1047). They suggest that energy varies along dimensions which mostly characterize differences between languages, speakers and recordings rather than phonation as such (Keating et al. 2011:1049).

2.9. THE PSYCHOACOUSTIC MODEL

Given that the focus of this study is on the acoustic measures described above and because of the aforementioned need to link these measures to perception in studies of voice quality, I will in this section review Kreiman, Gerratt, Garellek, Samlan and Zhang's (2014) relatively recently developed psychoacoustic model, which I will use in the interpretation of the acoustic results. This model is particularly useful since it seeks to delineate the link between physiology, acoustics and perception.

Kreiman et al. (2014:1-2) set out to develop a voice model which addresses what causes a given change in voice quality as well as what the impact of such a change on perception would be. The theory is thus specifically intended to describe the links between the production of voice and its perception (Kreiman et al. 2014:2). In developing this model, they used three main steps, firstly finding the link between perception and acoustics by explaining voice quality in terms of those acoustic measures which have the greatest perceptual validity and which in their combination can determine voice quality in full (Kreiman et al. 2014:2). Secondly, they intended to provide the link between both perception and acoustics with voice production by determining the physiological changes which produce those changes in the acoustic signal which are perceptually salient (Kreiman et al. 2014:2). The last step was to iterate the results from both the acoustic data and the physiological data until these two datasets aligned (Kreiman et al. 2014:2).

Kreiman et al. (2014:3) point out that, since voices appear to be processed as integral patterns rather than a sum of individual features, it is necessary in forming a unifying theory, to quantify the voice pattern in its entirety. In order to do this, Kreiman et al. (2014:3) used an analysis-by-synthesis approach. The reasoning behind this methodological choice rests on the fact that because the parameters of the acoustic synthesizer are combined in order to fully

recreate a voice pattern as it is perceived, those parameters included in synthesis can therefore be considered to constitute a psychoacoustic voice quality model of this pattern by means of parameters and thus provides an objective quantification of subjective speech percepts (Kreiman et al. 2014:3).

To help determine which of the acoustic parameters used varied the most across speakers (and thus would be the most likely to be perceptually salient), Kreiman et al. (2014:3) used a PCA (Principal Components Analysis) as detailed in Kreiman, Gerratt and Antoñanzas-Barroso (2007) for the spectra of 70 voices. The PCA indicated that most of the variance in the spectral shape of the source across speakers was accounted for by four factors, namely source spectral slope from 1.5 kHz through to 4 kHz (which comprised of two factors in their PCA analysis), the slope under 450 Hz and that above 4 kHz (Kreiman et al. 2014:3). When conducting a similar analysis using numerous acoustic measures, they also found that H1−H2 as well as H2−H4, in addition to the spectral slope overall, as well as high frequency noise were all important parameters (Kreiman et al 2014:3).

To assess the sufficiency of the model throughout the course of its development, Kreiman et al. (2010) copy-synthesized 700 voices based on both pathological and non-pathological speakers (Kreiman et al. 2014:4). Listeners were then asked to compare the natural voice samples with the synthesized tokens (Kreiman et al. 2014:4). The results suggested that greater detail was needed for source spectral modelling above H4 and so Kreiman, Garellek and Esposito (2011) and Kreiman and Gerratt (2011) replaced the H4−5 kHz parameter with two parameters, namely H4−2 kHz as well as 2 kHz−5 kHz (as described in the previous section) (Kreiman et al. 2014:4). After having done this, listeners were not consistently able to distinguish between the synthetic and natural voice tokens (Kreiman et al. 2014:4). Thus Kreiman et al. (2014:4) conclude that although constantly undergoing development, the model in its current form provides sufficient detail for the description of most pathological as well as most normal voice qualities. In order to determine the necessity of the individual parameters for inclusion in the model, a sequence of experiments were carried out in order to test the level of perceptual sensitivity of each parameter. First, sensitivity to the parameters was defined as the ratio of the just noticeable difference that speakers can detect for a given parameter to the overall

cross-speaker variability for the parameter in question (Kreiman and Gerratt 2010; Kreiman et al. 2014:4).

Following these experiments, (for more details, see Garellek, Samlan, Kreiman and Gerratt 2013 as well as Kreiman and Gerratt 2012), Kreiman et al. (2014:4) were able to conclude tentatively that the selected parameters meet the test of necessity and should thus be included in the psychoacoustic model. In the current model, the four parameters for the source include H1−H2, H4−2 kHz, H2−H4, and 2 kHz−5 kHz (Kreiman et al. 2014:4). Other components of the model include the inharmonic excitation of the source (which can be associated with harmonics-to-noise ratio parameters), time-varying source characteristics as well as the transfer function of the vocal tract (represented parametrically by formant frequencies) (Kreiman et al. 2014:4). There are certain claims implicit in the model, including that speakers should be able to control the parameters or alternatively the antecedent physiological events for the purpose of conveying information to listeners and secondly, that it should be possible to measure the parameters and that such measurements would provide an indication of those perceptible voice quality changes resulting from those aspects of production which speakers are easily able to manipulate (Kreiman et al. 2014:5).

Kreiman et al. (2014:5) claim that the psychoacoustic model finds support from evidence which supports these two main implicit assumptions of the model. For example, evidence from both high-speed imaging studies (see Kreiman et al. 2012) as well as studies investigating the linguistic use of voice quality for various languages supports the assumption that speakers are able to control either the parameters themselves or the antecedent physiological causes (Kreiman et al. 2014:5). Evidence for the second assumption, namely that those speakers should easily be able to manipulate those aspects of production which lead to perceptible voice quality changes (Kreiman et al. 2014:6), finds support from modelling studies such as that of Zhang et al. (2013). Zhang et al.'s (2013) research comprised a physical modelling study, in which both the acoustic effects of varying vocal fold stiffness as well as the perceptual effects were investigated. The fact that Kreiman et al.'s (2014) model explicitly links the production to the perception of voice quality makes it particularly useful for interpreting the results of the current study based on the results for the model parameters.

The second main implicit claim of Kreiman et al.'s (2014:6) psychoacoustic model is that speakers have the ability to easily control those factors which produce voice quality changes which are perceptible, which would in turn imply that by examining the perceptual consequences of physiological changes would enable the identification of those behavioural and mechanical manipulations which are most relevant perceptually.

As Kreiman et al. (2014:6) point out, while there are numerous studies which model laryngeal behaviour, few evaluate the perceptual effects of different model permutations. One notable exception to this, as Kreiman et al. (2014:6) note is Zhang et al.'s (2013) investigation of both perceptual as well as acoustic consequences of mismatches in left-right vocal fold stiffness in a mechanical, self-oscillating model of the vocal folds.

Zhang et al. (2013:453) set out to contribute towards an understanding across domains related to the causes and effects of the link between individual biomechanical vocal fold properties and voice quality by means of an investigation of both the perceptual as well as acoustic consequences of thyroarytenoid muscle activation. They focused specifically on the effect thyroarytenoid activity has on body-stiffening (Zhang et al. 2013:453).

As discussed by Kreiman et al. (2014:6-7), using their physical model, it was possible to explain changes in perception with reference to parameters of the psychoacoustic model, including both the noise-to-harmonics ratio as well as spectral slopes. For example, Zhang et al. (2013:457) found that for the condition in which the vocal folds were vibrating symmetrically, increases in body stiffness resulted in a reduction of H1−H2 and an increase in the number of harmonics observable at higher frequencies which would suggest a flattening of spectral slope overall. An increase in body stiffness was also found to be linked to a decrease in NHR (noise-to-harmonics ratio) which would be interpreted as a decrease in the production of noise (Zhang et al. 2013:457). Zhang et al. (2013:461) also found that changes which characterize the lower portion of the speech spectra were perceptually important to listeners, for example, for the measures H1−H2 and H4−2K.

Given that Kreiman et al.'s (2014) model directly links perception to acoustics, particularly for a number of the acoustic parameters included in this study, the model will be employed in providing plausible interpretations of the research findings of this study.

2.10. CONCLUSION

As is clear from the literature reviewed above, a number of sociolinguistic studies have included either as their main focus, or as a secondary focus, an acoustic analysis of voice quality. In the foregoing review, I have also provided a discussion of a number of key issues involved in studies of voice quality variation related to ethnicity in particular and a review of relevant literature pertaining to voice quality studies in the South African context. I have also provided a review of the acoustic measures used in this study and a summary of the psychoacoustic model which will be used in forming an interpretation of the results. In the following chapter, I will explain the research methodology employed in this study.

CHAPTER III: METHODOLOGY

3.1. INTRODUCTION

This chapter describes the research methodology employed in this study in order to meet the stated aims and objectives. The chapter provides details regarding the sampling and recording procedures, acoustic analysis using VoiceSauce (including a description of the different parameter estimates as implemented in VS), segmentation using PRAAT, the auditory analyses as well as the statistical analysis. Where appropriate, I have provided a condensed review of the relevant literature, in particular, with regard to the methods and techniques described in this chapter where these have been used by other researchers to address similar research questions.

3.2. SAMPLING

The sample was collected based on the hypothesis to be tested, namely that there are significant differences in voice quality between young Black middle-class English speakers (those who had attended ex-model C schools or private schools for most of their school careers) and White middle-class English speakers of the same educational background.

The sample for the research project was therefore composed of both monolingual White English speakers and Black speakers of an isiXhosa language background. The Black speakers included in the sample are therefore similar in terms of background to those included in Morreira's (2012) study, namely, those who have attended ex-model C schools (or private schools), since it has been observed that a difference in voice quality may be present for at least some in this group. I considered it necessary to select only those black participants who were of an isiXhosa language background. It was considered important to control for language background given the possibility that any voice quality differences may be linked to Bantu language transfer effects which could potentially obscure any trends in voice quality (in the event that voice quality effects may be different for those with different Bantu language backgrounds).

Only female speakers are included in the sample. The motivation for including only female speakers in the sample at this stage is twofold. Firstly, comparisons across sexes for certain voice quality measures are considered to be unreliable (Holmberg et al. 1988; Klatt and

Klatt 1990; Hillenbrand et al. 1994; Hanson 1997; Iseli and Alwan 2004; Iseli et al. 2007; Simpson 2009), which would necessitate the use of a sample of at least twice the size as that used in the current study (since each sex would need to be compared separately for these acoustic measures), which would be impractical at this stage.

Secondly, following a growing body of research, there is evidence that there is a greater crossing over by black female speakers (as opposed to black male speakers) towards the former white accent norms (Mesthrie 2010; Morreira 2012; Mesthrie 2017). The principal advantage of including only female speakers in the sample would therefore be that there is a greater likelihood of the black speakers being essentially identical to the white speakers in terms of vowel pronunciation and in particular, for the vowels selected as targets in the formal data elicitation materials. This would effectively reduce, if not eliminate any otherwise potentially large differences in vowel quality between the two ethnic groups being compared (as would have been anticipated if a comparison were to be made between traditional Black South African English and White South African English vowels). The sampling of only female subjects therefore also permits a more direct testing of the previously stated hypothesis, namely that there is an ethnolinguistic difference in voice quality even where segmental pronunciation is essentially identical.

3.2.1. Sample Collection

The first group of participants were recruited for the study, using a previously assembled list of contacts who had volunteered for earlier research studies but who were ineligible to take part in those studies. Subsequently, participants were recruited through the University of Cape Town student mailing list, as well as through campus-wide advertising on noticeboards at the University of Cape Town. The advertisements called for students who were fluent English-speakers and also those with an isiXhosa language background to take part.

Details of the schools which interested respondents attended were gathered through e-mail correspondence or telephonically, in order to help ensure that only those who attended ex-model C schools and private schools were selected of those who initially responded to the research invitation. In addition, potential participants who were regular smokers were excluded, since habitual inhalation of tobacco smoke is known to have an effect on voice quality (Laver 1975; Hollien, Hollien and DeJong 1997).

3.2.2. Sample Characteristics

As part of the interview segment of the recordings, certain biographical details were routinely elicited from the participants. The questions which were routinely asked during the elicitation are included in appendix B. All speakers recorded were between the ages of 18 and 22 in order to minimize the potential effect on the voice quality measures which would have been entailed by using a sample with a large age range (see Kreiman and Sidtis 2011 for a discussion of the effect of age on the voice). The sample consisted of 18 isiXhosa background research subjects (all of whom were black) and 18 white research subjects (thus 36 research subjects in total). Language background was controlled for. Language is reckoned to be more likely be of importance than region for voice quality, since there are no reports of distinct voice quality differences for different regions of South Africa and even for vowel quality variation, regionality is marginal for South African English compared to other varieties of English. All research subjects were English-dominant. A table of the dialect history of the speakers included in this study is supplied in the appendix. Most of the research subjects had grown up in the Western Cape (14 in total). Most of the white subjects had grown up in the Western Cape (10 subjects), while most of the black subjects had grown up in the Eastern Cape (8 subjects). That many of the Xhosa background subjects originally came from the Eastern Cape is unsurprising, given migration patterns within South Africa, (for a more detailed description of migration patterns and some of their implications for language use, see Deumert 2013). The rest of the research subjects came from cities and towns in Gauteng and KwaZulu-Natal.

3.3. RECORDING PROCEDURE

The recordings were conducted in a completely sound-proof recording studio. This venue was selected because of the sensitivity of a number of the measures to be used which would be adversely affected by extraneous and ambient noise from the environment.

The recordings were made using an AKG CM 311L head-mounted microphone connected to a Marantz PMD661MKII digital recorder at a sampling rate of 44.1kHz. After reading through the consent form, I provided the participants with a basic description of the tasks involved during the recording. Once they had given their consent, each subject was instructed to fit the microphone to her own head such that it was comfortable and I then checked the correct positioning of the microphone and made adjustments accordingly. This allowed for the

microphone to be placed at a constant distance of less than a centimeter from the speaker's mouth, as suggested as optimal by the operating instructions for the AKG CM 311L. The use of this microphone and recording equipment ensured that the audio input was recorded at a uniform distance from the source of the sound throughout the entire recording while still offering a high quality recording minimizing interference from any environmental noise. Participants were given the opportunity to ask questions about the tasks and the research in general before the commencement of recordings.

The participants were allowed to practice reading the first set of sentences in order to ensure that they understood what was required in each section. The participants were also instructed to speak at their normal pace and in a natural manner when reading through the list of sentences. No other control for intensity was used.

The list consisted of 110 different target words embedded in carrier sentences (of the form "Say uh,…again" as used for example in Iseli and Alwan 2004), each of which was repeated twice, as well as sentences with target words for mapping each speaker's vowel space, consisting of 17 sentences (see appendix A). Thus the total number of sentences for each speaker was 237. The total number of tokens included in the sentence data was 7772, with an average of 216 tokens per speaker. These sentences were presented on a series of PowerPoint slides on a computer screen display in the adjacent room. The screen was positioned behind a glass panel, so that it was visible to participants. I controlled the rate at which the slides were shown, such that list intonation (Kreiman and Sidtis 2011:288) was avoided. Following the reading of the sentences, each participant was interviewed. The style of interview was similar to the well-known Labovian sociolinguistic interviews, during which fairly detailed information about topics such as language background and schooling history were collected. These interviews also included a modified version of Labov's 'danger of death' question, in the form of a question about the participants' experiences of crime, designed to elicit more spontaneous speech data. In addition to providing detailed information on each participant's background, which would enable the researcher to potentially identify relevant social variables and to identify and exclude outliers where necessary, the interviews also served to provide the core data of this study for the acoustic analysis of voice quality as well as a possible comparison with the formal style speech data as provided in the read sentences, which would allow for a comparison of the consistency of any

differences observed across contextual styles. I monitored the recording level throughout, using the level indicator of the Marantz PMD661MKII in order to prevent clipping.

3.4. ACOUSTIC DATA ANALYSIS USING VOICESAUCE (VS)

As mentioned in earlier chapters, VoiceSauce (VS henceforth) is a program specifically designed for the analysis of voice quality by extracting from the acoustic signal those parameter estimates relevant for the analysis of differences in phonation as reviewed in the previous chapter (such as H1−H2, for example). A full description of the capabilities of VS is provided in the online documentation (Vicenek, Lin, Keating and Shue 2017).

3.4.1. Initial Data Preparation

In order to permit an analysis using the parameter estimates provided by VS to be carried out, the recorded .wav files were first loaded into PRAAT for splicing into smaller .wav files. This was done because VS works more efficiently with smaller files (Shue, p.c. and Vicenik, Lin, Keating and Shue 2016).

It is recommended that the files be downsampled to 16 kHz (Shue, p.c.). This procedure also makes the results more directly comparable with those of Iseli, Shue and Alwan (2007), who also used a sampling rate of 16 kHz. VS offers the option 'Process using 16kHz,' which downsamples the files. For this reason, I considered it unnecessary to downsample the smaller files first (using PRAAT to do this, for example), since it was done automatically in VS using the downsample option before data processing.

Each of these files were subsequently converted to mono .wav files, on Shue's (p.c.) recommendation, since VS only works with the first channel of multi-channel wav files. The conversion was done using PRAAT.

The files were then segmented and annotated using the recommendations for segmentation in VS (Vicenik, Lin, Keating and Shue 2017).

3.4.2. Annotation in PRAAT using PRAAT Text Grids

PRAAT Text Grids were used to annotate each of the files. Orthographic transcription of the read sentences and interviews was done by myself as the principal researcher and two research assistants, who were trained by me to transcribe these files using PRAAT Text Grids. I subsequently checked each of the transcribed Text Grids for transcription errors and corrected these accordingly. For the sentence data, where participants were instructed to repeat a given utterance due either to a false start, misreading or any other kind of interruption, only the second reading was transcribed. Only the interviewee's speech was transcribed. Each transcription was converted into a tab-delimited text file in a format required for processing in FAVE (described below). A PRAAT script written by Ingrid Rosenfelder, was used to convert the PRAAT transcriptions into the required tab-delimited text files.

The next step in the alignment process is to conduct an 'unknown words check,' that is, to check that the words in the orthographic transcription are present in the CMU (the Carnegie Mellon University) pronouncing dictionary, a pronunciation dictionary which is both open-source and machine-readable, originally designed for the description of North American English (CMUdict 2015). The aligner uses this dictionary such that the unknown words can be manually added for the purpose of automatic alignment. The list of unknown words is generated by selecting certain options available on the online alignment suite. The file containing the words which have no corresponding entry in the CMU pronouncing dictionary are then sent back to the user after being written to file.

The transcriptions for these out-of-dictionary words were then supplied, using the ARPAbet (developed by the Advanced Research Projects Agency) conventions, a system used for phonetic transcription representing each phoneme of General American English using distinct ASCII character sequences. The input transcription file (using the ARPAbet conventions) was then re-uploaded via the website interface by selecting the 'import dictionary transcriptions' option. This was uploaded along with the original transcription file as well as the associated audio file for alignment.

The output of this process was a TextGrid sent by the site as an e-mail attachment for each set of files uploaded. I subsequently checked the aligned files and ensured that the vowel

tokens for the three vowel categories (STRUT, THOUGHT and FLEECE) were correctly labeled.

3.4.3. Automatic Segmentation using FAVE (Forced Alignment and Vowel Token Segment Extraction)

Segmentation was done automatically using the online version of FAVE-align (Rosenfelder, Fruehwald, Evanini and Jiahong 2011). This is a program suite for forced alignment developed by Jiahong Yuan and Mark Liberman at the Linguistics Lab of the University of Pennsylvania as an adapted version and is based on the original Penn Phonetics Lab Forced Aligner (or P2FA), but is more specifically tailored for use in segmenting transcribed interviews.

FAVE was chosen for segmentation purposes rather than opting for manual segmentation for two main reasons. Firstly, motivated by purely practical considerations, given the fairly large amount of speech data to be segmented prior to analysis, it was found to be more economical with respect to time to use automatic segmentation. Secondly, automatic segmentation was considered to constitute a more objective means of segmentation that would reduce the possibility of human error inevitably necessitated by a process of purely manual segmentation (since while manual checking was involved, relatively few tokens required re-alignment as discussed below), allowing for the results obtained using this method to be more easily replicated in future research than would otherwise have been possible.

The primary goal in segmentation, as far as the acoustic component of the analysis is concerned, was the efficient extraction of only voiced portions of each of the vowel tokens to be analyzed, such that the maximum number of parameter estimates provided by VS could be reliably used, since a number of these estimates are only reliable for voiced speech segments.

A general overview of how FAVE-align achieves automatic segmentation is provided in the following section. Readers interested in more detailed information regarding the workings of FAVE-align are encouraged to read the information available on the website (Rosenfelder et al. 2011).

3.4.4. Segmentation Checking and Extraction using PRAAT

Once each PRAAT TextGrid file had been automatically aligned using FAVE-align (Rosenfelder et al. 2011), with the help of a trained research assistant, I manually checked the segmentation in order to ensure that there was no misalignment (such as occasionally occurred with tokens

preceded by a nasal or liquid, for example, where some of the nasal or liquid portion was included in the vocalic segment) and that only voiced, reasonably stable segments had been included as vowel segments. In cases of misalignment, the boundaries of the TextGrid were accordingly shifted based on an auditory and visual inspection of the spectrogram, waveform and associated TextGrid in PRAAT. While FAVE-align may occasionally make such errors in alignment, for my data, such errors occurred relatively infrequently and therefore the aforementioned advantages of using FAVE-align in terms of saving time and in terms of affording increased replicability were considered to outweigh any disadvantages potentially entailed by its use.

In line with previous studies investigating voice quality variation by means of the acoustic correlates used in the current research (Hanson 1997; Iseli and Alwan 2004; Iseli, Shue and Alwan 2007) and as recommended by Esling and Edmonson (2011), tokens were segmented such that they contained, so far as possible, steady state portions of the vowels in question to allow for more reliable measurements to be taken.

Once the automatic segmentation had been checked and the TextGrid boundaries realigned where necessary, PRAAT was used in order to extract each labeled vowel segment. These extracted files were then opened in VS for acoustic analysis.

3.5. VS PARAMETER DESCRIPTIONS

In this section, I provide details of how the various parameter estimates described in the previous chapter are implemented in VS.

3.5.1. Harmonic Differential Measures

VS provides a number of harmonic differential measures. These measures derive their utility for the acoustic study of voice quality variation from observations that different voice qualities are expected to exhibit certain predictable effects on the harmonic structure of the speech signal, as described in the previous chapter. In this study, both corrected and uncorrected measures were computed. Although the focus in the results chapter is on the results for the corrected measures, basic summaries of the findings for the uncorrected measures are also provided. This is firstly to

allow for a comparison where relevant with earlier studies where only the uncorrected measures were used and also because this information may be useful for future voice quality research.

3.5.1.1. H1–H2[45]

For this measure, as well as the other harmonic differential measures as implemented in VS, the magnitudes of the harmonic spectra are calculated pitch-synchronously using a 3-cycle window (Shue et al. 2011). This allows for a great deal of the spectral variability found for spectra computed using fixed time windows to be eliminated (Shue et al. 2011). VS finds the harmonic magnitudes using an algorithm which conducts a maximum search around the locations of the spectra as estimated by fundamental frequency, with the search range being restricted to 10 percent of the estimated value for the fundamental frequency (Shue et al. 2011). The advantage of this method, as pointed out by Shue et al. (2011) is that while it is the equivalent of using an extremely long Fast Fourier Transform window, it enables more accurate measurements to be made without the need to rely on extensive FFT calculations.

Vicenik, Lin, Keating and Shue (2017) note that the reliability of the H1−H2 measure and other such harmonic differential measures is dependent on the correct estimation of the parameters which form part of these measures. Therefore for all the harmonic measures calculated by VS, it is necessary to check that the f_0 and formant frequency tracking by the program are correct too (Vicenik, Lin, Keating and Shue 2017). This was done in each case for each sample for each speaker, by checking the individual files using PRAAT.

3.5.1.2. H2–H4

The corrected measure in VS uses the first and second formant values as estimated by VS for the correction of the effects of formants using the formula developed by Iseli and Alwan (2004)[46]

[45] Note that for most of the harmonic measures provided by VS, corresponding corrected measures are also provided, which correct for the effects that formants and their bandwidths have on the harmonic amplitudes. Thus, for example, VS offers both H1−H2, as well as the corrected measure H1*−H2*, where by convention, the asterisks indicate that the respective harmonic amplitudes have been corrected for the influence of formants and formant bandwidths. In this study, both corrected and uncorrected measures were computed, but the analysis focuses on the corrected measures. This applies to all of the spectral tilt measures described in this section.

[46] Iseli and Alwan's (2004) formula for calculating corrected harmonic magnitudes is as follows:

and for the bandwidths using Hawks and Miller's (1995) formula[47] to correct for the effect of formant bandwidths for the first and second formants. As observed by Vicenik, Lin, Keating and Shue (2016), if the fourth harmonic (H4) happens to be found at a formant frequency, there may be a large error in the corrected measure H4* when using Hawks and Miller's (1995) formula. While measured bandwidths as provided by VS may potentially reduce such errors, because these are not guaranteed to be correct (Keating, p.c.) either, I ultimately selected the Hawks and Miller (1995) formula for the bandwidth correction for H4 in this study.

3.5.1.3. H1–A1

This measure (and its corrected counterpart) as implemented in VS involves the subtraction of the amplitude of the closest harmonic to F1 from the first harmonic amplitude. VS provides the corrected measure (H1*−A1*) by correcting for each frame using the formant frequencies as measured by the program and uses the formula-derived bandwidths (as mentioned above) of

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

where N represents the number of formants, $H^*(\omega)$ signifies the corrected harmonic magnitude at given frequency ω, and $H(\omega)$ represents the original signal spectral magnitude (measured in decibels) at frequency ω. For the full derivation of this formula, see Iseli and Alwan (2004).

[47] Hawks and Miller's (1995) formula for the calculation of formant bandwidths is as follows:

$$F_b = S * \left(k + (x_1 * F_c) + (x_2 * F_c^2) + (x_3 * F_c^3) + (x_4 * F_c^4) + (x_5 * F_c^5)\right) \quad (2)$$

where F_b represents the calculated formant bandwidth, F_c is the "formant center frequency" and S is a scalar (for accommodating the expected wider bandwidths for speech produced by female speakers) derived from the following formula:

$$S = 1 + 0.25\left(\frac{F0 - 132}{88}\right) \quad (3)$$

(Hawks and Miller 1995). For the full derivation of this formula as well as the exact values for k and x as found in equation (2) (which differ according to whether formant centre frequencies are above or below 500 Hz), see Hawks and Miller (1995).

those frequencies for bandwidth correction (Shue et al. 2011). Finally, VS then smooths the measures by means of a moving average filter[48], 20 samples in length (Shue et al. 2011).

3.5.1.4. H1–A2

This harmonic differential measure as implemented in VS involves the subtraction of the amplitude of the strongest harmonic in the second formant region from the amplitude of the first harmonic. The corresponding corrected measure (H1*−A2*) uses the values for the first two formants and the values for the first two formant bandwidths calculated by the formula (that of Hawks and Miller 1995) to correct for the effects of formants and their bandwidths.

3.5.1.5. H1–A3

The harmonic differential measure H1−A3 as implemented in VS involves the subtraction of the amplitude of the strongest harmonic in the F3 region from the amplitude of the first harmonic. The corresponding corrected measure (H1*−A3*) as implemented in VS corrects for the effects of formants and their bandwidths on the harmonic amplitudes using the formula developed by Iseli et al. (2006) and using the formant values as estimated by VS for the first two formants for the correction of H1 to H1* and the first three formants for the correction of A3 to A3*. A3 also uses the bandwidth values for the first three formants as calculated from Hawks and Miller's (1995) formula for the bandwidth correction.

3.5.1.6. H4–H2K

This measure as implemented in VS involves the subtraction of the amplitude of the strongest harmonic closest to 2000Hz from the amplitude of the fourth harmonic. The corrected measure

[48] This is one of the most common filters used in digital signal processing and is optimal for the reduction of random noise with simultaneous retention of sharpness of the step response (Smith 2003:277). It works by averaging a number of adjacent points from the input signal "to produce each point in the output signal" (Smith 2003:277). It is known as a 'moving average filter' because it first looks at a sequence of samples through a window of a particular size (in this case, the window size is 20 samples), computes an average value and then moves to compute the following average (Vetterli, Kovačević and Goyal 2014:201). These average values form the points of the output signal produced by the filter. This type of filter performs exceptionally well at smoothing data and is therefore traditionally used as a smoothing filter (Roberts 2008:525; Smith 2003:280).

uses the first two formants and their bandwidths for the correction of H4 to H4*, but relies on the first three formants for the H2K* corrected measure (Vicenik, Lin, Keating and Shue 2017).

3.5.1.7. H2K−5K

The calculation of this measure involves the subtraction of the highest amplitude harmonic in the 5000Hz region from the highest amplitude harmonic in the 2000Hz region.

In VS, only the harmonic with the highest amplitude in the 2000Hz region is corrected for the effects of formants and their bandwidths, using the first three formant values and the bandwidth estimates provided by VS, since it has been observed that there is a significant increase in the inaccuracy of the estimation of formants the higher the formant values (Vicenik, Lin, Keating and Shue 2016). For this reason, the developers of VS have not included a formant corrected measure for the value of 5K. H2K is corrected to H2K* as described above.

3.5.2. Measures of Signal Aperiodicity and Noise in VS

3.5.2.1. HNR (Harmonics-to-Noise Ratio)

The HNR measures as implemented in VS are derived from the algorithm for estimating HNR developed by de Krom (1993), as reviewed in the previous chapter (Shue et al. 2011). In VS, this measure uses a window of variable length which by default, is equal in length to 5 pitch periods (Shue et al. 2011). The process involves the liftering of the cepstral pitch component and then making a comparison of the harmonic energy with that of the noise floor (Shue et al. 2011). This ratio is calculated for four frequency bands, namely between 0 and 500Hz, between 0 and 1500Hz, between 0 and 2500Hz and between 0 and 3500Hz (Shue et al. 2011; Vicenik, Lin, Keating and Shue 2017).

3.5.2.2. Cepstral Peak Prominence (CPP)

The CPP parameter estimate calculated in VS employs the algorithm first developed by Hillenbrand et al. (1994) and uses as a default, a window varying in length but equal to 5 pitch periods (Shue et al. 2011). The data are subsequently multiplied using a Hamming window and are then transformed into the domain of the real cepstrum (Shue et al. 2011). A maximum quefrency search is performed by VS around the pitch period to arrive at the CPP value (Shue et

al. 2011). The cepstral peak then undergoes normalization to a linear regression line calculated between the maximum quefrency and 1ms (Shue et al. 2011). Note that an important difference between CPP and HNR measures in VS is that they cover different frequency ranges.

3.5.2.3. Subharmonic-to-Harmonic Ratio (SHR)

The SHR (or Subharmonic-to-Harmonic Ratio) as implemented in VS, follows the measure which Sun (2002) proposed and provides a quantification of the ratio between the harmonic amplitudes and the subharmonic amplitudes. VS implements Sun's (2002) code as well as Sun's (2002) algorithm for this measure (Shue et al. 2011). VS derives the SHR measure by summing the amplitudes of the subharmonics and harmonics using spectral shifting within the log domain (Shue et al. 2011).

3.5.3. Energy (Root Mean Square Energy)

The energy measure used in VS is that of RMS (or root mean square) energy, which is calculated over a window of variable length but equal in each case to 5 pitch periods and it is computed for every frame (Shue et al. 2011). The use of the variable window is desirable because it in essence normalizes this measure in order to reduce the correlation between f_0 and the measure itself (Shue et al. 2011).

A summary of the measures used in this study and how they relate to voice quality (based partly on Garellek 2016, as well as the foregoing review) is provided in table 3.1 below.

Types of Measure	Measure	Basic description	Relevance for Voice Quality Analysis
Measures of Spectral Slope	H1*−H2*	First harmonic amplitude minus second harmonic amplitude	Open quotient, degree of glottal constriction
	H2*−H4*	Second harmonic amplitude minus fourth harmonic amplitude	Stiffness and thickness of the vocal folds; breathiness perception, characterizing creaky vowels, speaker sex identification
	H4*−2 kHz*	Fourth harmonic amplitude minus the amplitude of the strongest harmonic at 2000 Hz	Vocal fold thickness; involved in contrastive breathiness perception
	H2 kHz*−H5 kHz	Amplitude of the strongest harmonic at 2000 Hz minus the amplitude of the strongest harmonic at 5000 Hz	Vocal fold stiffness; involved in the perception of breathiness for some languages
Measures of Spectral Slope (Formant-based)	H1*−A1*	First harmonic amplitude minus the amplitude of the strongest harmonic in the F1 region	Breathiness related to posterior glottal gap
	H1*−A2*	First harmonic amplitude minus the amplitude of the strongest harmonic in the F2 region	Overall spectral tilt and abruptness of closure, Distinguishes breathy from modal phonation in some languages
	H1*−A3*	First harmonic amplitude minus the amplitude of the strongest harmonic in the F3 region	Same as above; glottal pulse skewness
Noise Measures	CPP	Measure of the relative strength of the noise component to harmonics	Distinguishes modal from non-modal; aperiodicity; plays a role in the perception of breathiness
	HNR	Measure of the relative strength of the noise component to harmonics	Distinguishes modal from non-modal; associated with added noise or voicing irregularity
	SHR	Measure of the strength of subharmonics	Multiple pulsing; used in identifying creak sub-types

Table 3.1: Summary of measures used in this study and their relevance for voice quality.

3.6. GENERAL VS SETTINGS

I selected the VS data output option to take measurements at 1ms intervals (of all values measured at the 10ms frame shift rate) for each labelled segment. While VS does provide the user with the option of taking measurements at greater intervals, as well as of computing

averages over specified subsegments, the way I have operationalized voice quality in this study (based on Esling and Edmonson 2011), is that it is a phenomenon which is essentially present more or less the entire time one speaks, at least for voiced speech. For this reason, selecting the largest number of measurements provided by the program as possible, rather than restricting the analysis to a small number of measurements for each segment would more accurately capture any long-term differences in voice quality. Since there was no a priori motivation for selecting a particular number of subsegments to average over and because averaging operations could easily be performed using the statistical analysis program R, post hoc, provided the full number of possible measurements had already been taken, I made the decision to take the full number of measurements offered as an option in VS for the sentence data. The definition of voice quality adopted in this thesis allows it to be conceptualized as a type of average taken over many segments for a given speaker. Therefore, in line with this definition and in order to make the analysis less computationally expensive, more time efficient and to allow for a greater variety of statistical analysis methods to be used as needed, I averaged measures over 1 subsegment for each vowel segment included in the analysis to obtain the VS output.

In order to make my results more directly comparable with those of researchers using PRAAT algorithms (for example, Shue et al. 2011 and Szakay and Torgersen 2015), I used the PRAAT option for pitch and formant tracking using the default settings. Adjustments of these settings for each segment were avoided since any such adjustments would be likely to introduce an undesirable level of subjectivity and would decrease the reliability of the measurements. Iseli et al. (2006) used a window shift rate of 10 ms, as did Iseli et al. (2007). Thus, to make the results comparable therefore, the same options have been selected in conducting the analysis in the present study.

3.7. ACOUSTIC MEASUREMENT OF FORMANT FREQUENCIES USING THE PRAAT ALGORITHM IN VS

VS provides estimates of formant frequencies as well as formant bandwidths using several options, including PRAAT, and SNACK sound toolkit. Since PRAAT was used for the estimation of the voice quality parameters, this option was, for the sake of consistency, likewise

selected for the estimation of formant frequencies. The differences in formant frequencies were therefore also subjected to statistical analysis using R, following the procedures for statistical analysis outlined later in this chapter.

3.8. AUDITORY ANALYSIS

Following the acoustic analysis as described earlier in this chapter, an auditory analysis was also performed on the sentence data. I considered this necessary, in order to provide some insight into the potential perceptual relevance of any observed differences in the patterning of the acoustic measures. The auditory analysis was also used in order to provide some insight into the frequency of use of phonation types (rather than long-term voice quality) for the two ethnolinguistic groups. As noted in the final chapter, in future research it will ultimately be desirable to test the perceptual relevance of such differences to South African listeners.

Careful consideration was given to how the auditory analysis should be carried out as there were a number of options available, as illustrated in the discussion of the different methods used by researchers investigating similar phenomena to follow. In this section I will review and evaluate these methods of auditory analysis with a particular focus on those studies which have used them in investigations of similar phenomena as those of the current study and with similar aims.

A number of auditory protocols and techniques have been used in the past in providing a means of quantifying and evaluating sociolinguistic voice quality variation, some of which have already been reviewed in the previous chapter.

For example, Esling's (1978) investigation of voice quality in Edinburgh involved several stages of an auditory analysis, in addition to an acoustic analysis using laryngographic waveforms. The auditory analysis in this case consisted of judgements of voice quality by Esling (1978) himself, a repeated analysis by Esling two years later, an analysis by trained phonetician judges and finally an auditory analysis by judges (in this case, five postgraduate phonetics students) who had been explicitly trained by John Laver in the description of voice quality according to Laver's (1975) descriptive protocol. The auditory judgement by the phoneticians

was included in order to provide an assessment of the consistency of Esling's (1978:128) own analysis with the state of the art in voice quality description at the time.

Esling (1978:128) used three main procedures in describing articulatory setting auditorily (including a description of phonatory settings). Firstly, the main features characterizing a particular informant's voice quality were identified, followed by the use of Laver's (1975) labeling system for voice quality features to accordingly label those features and finally, in order to discover which features were most frequent in each particular group (according to predefined social divisions of class and region), the judgements thus obtained were compared across the different groups for the whole sample (Esling 1978).

The labeling system developed by Laver (1975) allows for voices to be rated on any one feature on a scale from 1 to 3 with the slight presence of a particular feature being assigned the value of 1, an assignment of the value of 3 for where one feature was found to be extreme for a particular voice and a moderately noticeable feature being assigned a value of 2 (Esling 1978). If a feature is completely absent, it was given a rating of 0 (Esling 1978). Overall tension settings are specified in this system as either tense, lax or neutral (Esling 1978).

Esling (1978) points out that the reliability of such a technique which makes use of auditory judgements depends on the training of the phonetician making the judgements in being able to do so more or less uniformly and reliably for an entire speech sample and in terms of the commonly established definition for each label.

A similar type of auditory analysis was used by Stuart-Smith (1999) in her investigation of the voice quality features of the Glasgow accent. In order to proceed with this analysis, it was necessary to assume a fairly uncomplicated relationship between the articulatory settings and their auditory impression (Stuart-Smith 1999). Stuart-Smith (1999) made use of an adapted version of Laver, Wirz, Mackenzie and Hiller's (1981), Mackenzie Beck's (1988) and Laver's (1991) Voice Profile Analysis (or VPA) protocol, which, according to Stuart-Smith (1999) is particularly well-suited for voice quality transcription. For each of the speakers included in her sample, a 'VPA profile' was separately established for their read and conversational speech respectively (Stuart-Smith 1999). These speakers were transcribed in a random order such that their social group was not reflected in the order in which they were transcribed (Stuart-Smith

1999). The VPA's for each speaker were both quantitatively analyzed (in the form of descriptive statistics for shared features within social groups) as well as qualitatively analyzed (by writing out a verbal descriptive summary of each participant's VPA profile) (Stuart-Smith 1999). These descriptions were then pooled into groups such that shared features between different class, age and gender groups would become apparent (Stuart-Smith 1999). Subsequently, the group values were conflated in such a way as to reveal the shared features of larger social groupings (Stuart-Smith 1999).

Henton and Bladon (1988) in their study on creak and its function as a sociophonetic marker in two dialects of British English, also made use of a primarily auditory analysis. These researchers motivate their choice for using the auditory analysis (using acoustic analysis only as supporting evidence for their claims based on the auditory analysis) by stating that the perceptual effect of creak is easily auditorily identifiable (as a sound similar to 'popping corn') and is 'well-known' to both trained and untrained listeners (Henton and Bladon 1988: 9-10). Therefore these researchers conclude that the use of an auditory criterion for establishing the presence of creak in their data would be sufficient (Henton and Bladon 1988).

Henton and Bladon (1988) listened to and scored their sound recordings three times, scoring each syllable for creak for those syllables which were judged to contain it, using the auditory criterion of separately resolvable glottal vibrations. They then listened to the corpus a fourth time in order to reach agreement on scoring, stating that there was not much difficulty in reaching consensus between the two of them since 'the presence of creak is usually rather auditorily apparent' (Henton and Bladon 1988:16).

Henton and Bladon (1988) then used spectrograms which were selectively sampled rather than systematically measured for their entire corpus, for the purposes of illustrating the differences in acoustic terms between those syllables judged to be creaky and those judged to be modal.

More recently, in his study on the use of phonation as a stylistic variable, Podesva (2007) made use of auditory measures in identifying falsetto. In this case, an auditory analysis was used primarily because there was no single acoustic measure which could be used to provide an adequate differentiation between modal and falsetto voice (Podesva 2007).

Thus, in this case in particular, while the identification of the phonatory qualities of interest using acoustic measures was problematic, perceptual criteria could reliably distinguish between them and therefore Podesva (2007) made use of an auditory analysis which was supported by an acoustic analysis after the occurrence of falsetto had been auditorily determined. Podesva (2007) checked the accuracy of his identifications of falsetto with an informant who had undergone some training in singing and phonetics to check for which syllables in the data falsetto was present. The judgements of the informant matched those of Podesva (2007) 94% of the time, indicating a high degree of accuracy. The pattern for where there were differences in judgement (namely, cases where the informant labeled as 'modal voiced' tokens which Podesva had labeled 'falsetto') suggested that they both made use of similar judgement criteria for the decision, differing only in terms of how conservative the judgements were, with the informant being slightly more conservative in this regard (Podesva 2007: 484).

Podesva (2007) generated wide-band spectrograms, pitch tracks and waveforms to aid in the supporting instrumental analysis which made use of the following measures: falsetto and creaky voice duration, maximum and minimum f_0 and time of maximum and minimum f_0. Formulae were then used for deriving several other variables for the analysis such as the range of f_0 and rate of f_0 change (Podesva 2007).

Yuasa (2010) also made use of a similar auditory analysis in addition to using acoustic measures. Yuasa (2010) was interested in the use of creaky voice as 'a new feminine voice quality' in America. Yuasa (2010) used recordings of 10 minute conversations with her participants.

After the transcription of the recorded conversations, Yuasa (2010) counted the number of words which contained creaky voice. The lowest total number of words in the recording of any of the participants was taken as the total number of words to be extracted from each recording for analysis based on the presence of creaky voice (Yuasa 2010). This number was 401 and therefore this number of words was randomly selected for each recording (Yuasa 2010). Yuasa (2010) further assumed that a length of 4 minutes and 20 seconds (the maximum length of

time it takes to utter 401 words as selected) would be sufficient for the analysis of creaky voice, since creaky voice would not expected to be present for every word. Yuasa (2010) is in agreement with Henton and Bladon (1988) that creak is easily auditorily recognizable, based on their description of creak as well as that of Hollien et al. (1966: 246) who defined it as 'a train of discrete laryngeal excitations, or pulses of low frequency' (Yuasa 2010: 323-324).

Wherever Yuasa (2010) found creaky voice to be present for a particular word extracted from the speech of one of her participants, PRAAT 4.1. was used for the examination of both spectrograms and waveforms in order to look for the acoustic attributes for this phonation type as identified in the relevant literature, namely irregular glottal pulses and vertical striations spaced well apart in the spectrogram display as indicative of vibrations of low-frequency. Yuasa (2010) further affirmed the validity of using an auditory analysis in that this analysis corresponded in every case to the well-spaced vertical striations on the spectrogram for every token judged to contain creaky voice auditorily. Each instance of creaky voice derived from the auditory criteria confirmed by the acoustic analysis (as described above) was coded as one instance of creaky voice to allow for a frequency count (Yuasa 2010). Finally, Yuasa (2010:325) ran t-tests to test for significant differences for the informant group means.

Podesva (2013) in his investigation of gender and the social meaning of non-modal phonation types also made use of an auditory analysis of voice quality. The procedure used by Podesva (2013: 429) involved the individual coding of syllables for one of six phonation types, namely falsetto, creak, breathy voice, harsh voice and whispery voice. For the difference between breathy and whispery voice (which while being articulatorily distinct, are not necessarily perceptually distinct), Podesva (2013) operationalized the former as being voiced, while whispery voice was defined as being voiceless. For each auditory label, supporting acoustic evidence was provided by using spectrograms, pitch tracks and waveforms, using PRAAT for repeated listening to the isolated syllables and a second researcher double checked the coding (Podesva 2013). In this study, for every intonation phrase a percentage was given indicating the percentage value of each phonation type present in that intonation phrase. As far as the statistical analysis conducted on these data is concerned, Podesva (2013) used the same technique as is used in the current study, namely linear mixed effects regression.

Goodine and Johns (2014) likewise used an auditory coding for creaky voice, supported by subsequent acoustic measurements, in their study of the use of vocal fry by female Canadian English speakers.

Given the foregoing discussion, I decided that it would be appropriate, necessary and in line with current best practice in the study of voice quality variation, to conduct an auditory analysis using similar methods to those which have been outlined, particularly in more recent studies. Since the initial acoustic analysis had suggested possible differences in the prevalence of creak within each of the two groups of speakers in this study, I decided to further investigate the use of creak by the participants by means of auditory coding of creak. This form of analysis has as one of its advantages, as pointed out by other researchers (for example, Henton and Bladon 1988 and Yuasa 2010), that creak in particular, is easily identified by auditory means. Following Yuasa (2010) and Podesva (2013), I coded extracted tokens according to the auditory presence or absence of creak by first identifying them as such impressionistically and then checking the waveforms and spectrograms for these tokens for the acoustic properties of creak by means of visual inspection.

Recently, several advances have been made in the study of non-modal phonation types, particularly for different types of creak. Keating et al. (2015) have for example, identified six different types of creak and have exemplified the spectral correlates and waveform displays of these different types of creak, providing descriptions of general trends in terms of VS parameter measures for each. This is extremely helpful for researchers investigating creak, particularly for those making use of acoustic methods to do so. For each of the six types of creak, examples of waveforms are given, which can be useful in identifying the types of creak from the waveforms in addition to using the other spectral measures (Keating et al. 2015). For example, the regular alternation between weak and strong pulses as seen in a given waveform is indicative of multiply pulsed voice, one of the several types of creak described (Keating et al. 2015).

Likewise, if I identified any particular extracted token to be breathy, for example, supporting evidence was sought in the waveform and spectrogram for that token and when confirmed, the token in question would be coded as breathy. For this exploratory sentence data, there were tokens from each vowel category (there were three lexical sets used, namely STRUT, FLEECE and THOUGHT) extracted from the key words in each of the read sentences. For each

of these tokens, VoiceSauce extracted measurements every millisecond. This resulted in a total of 2977 measurement points for FLEECE, 2622 measurement points for THOUGHT and 2174 measurement points for STRUT, as detailed in table 4.1 on page 123 in the following chapter. This was done in order to capture the frequency of phonation types more accurately, which was ultimately one of the more important aims of the auditory analysis of the exploratory sentence data. Since several phonation types can be present at different times during a vowel, averaging out these differences for each vowel, would fail to accurately capture the frequency of phonation types. For the sentence data, the STRUT, FLEECE and THOUGHT lexical sets were chosen as the vowel tokens in the target words. These vowels of different qualities (one low, one high, one back) were used in order to examine the effect of vowel quality on the acoustic measures provided by VoiceSauce (VS). For speakers who have attended ex-model C schools, we would expect that vowel quality differences here would be negligible.

While I have coded the entire dataset for the sentence data for all speakers in this way, thus providing auditory data for the entire sample which can be statistically compared with the acoustic measures extracted using VS, for the initial auditory analysis, two speakers were selected from each group as those most clearly exemplifying the overall differences in voice quality judged to impressionistically distinguish the two groups. These two speakers can therefore be considered to exemplify the general distinctions in voice quality between the two groups which may be related either exclusively, or primarily, to a difference in the use of creak (or use of different types of creak), based on my overall impression and the preliminary acoustic analysis using VS. The data for these two speakers were also auditorily coded by a trained research assistant as described below in order to assess the usefulness of the coding procedure for South African English when used by relatively inexperienced coders.

Therefore, for this initial auditory analysis, the speaker selected from the white subsample was selected on the basis that her voice quality most clearly exemplified the norms (for the use of creak, in particular) for her ethnic group, while the black speaker for this analysis was selected on the basis that her voice quality was typical of the norms for her ethnic group, specifically in terms of the use of creak and greater use of breathy voice.

The data from these speakers were then coded as described above and were additionally checked and coded by the research assistant. I provided the assistant with guidelines and

references for examples of different phonation types, particularly Esling and Edmonson's (2011) online resources for most phonation types and Keating et al's (2015) guidelines for the identification of different types of creaky voice. The speech segments extracted from the read sentences were provided to the assistant whom I provided training to in the auditory analysis of phonation types by going through several examples. PRAAT was used to play the files and examine the waveforms and spectrograms. The procedure was to play and listen to the sample in PRAAT and if any non-modal phonation type was suspected based on the auditory impression, then the waveform and spectrogram display was opened and examined for any evidence of a particular non-modal phonation type. If I found any evidence based on a visual inspection of the waveform and spectrographic display that the segment was an example of a particular non-modal phonation type, that segment would be coded with the label for the corresponding non-modal phonation type, for example, as "breathy," "vocal fry" etc.[49] In all other cases, the segments were coded as modal. Whenever a segment was auditorily judged to contain some form of creak but there was no spectrographic or waveform evidence of any particular subtype of creak, the segment was coded as containing prototypical creak. In my own coding of the data I found very few examples of either aperiodic voice as described in Keating et al. (2015) and no unambiguous examples of non-constricted creak[50]. Segments which were clearly non-modal but could not be classified according to any of the categories provided were classified as 'unknown.' Very few segments were classified as 'unknown.' While it was relatively straightforward to distinguish between different phonation types containing creak and to distinguish creaky phonation types from those not containing any creak, it was considerably more difficult to identify breathy segments by ear.

This is expected, since, as Gerrat and Kreiman (2001:377) point out, due to the fact that breathiness is a continuous variable and that a continuum exists between modal voice and breathy voice, it is in practice difficult to distinguish between the two phonation types. Gerratt and Kreiman (2001:377) point out that a number of listeners are unable to reach agreement on whether a particular voice is breathy or not when assessing breathiness auditorily and this applies equally to trained as well as untrained judges.

[49] Descriptions of auditorily identified phonation types are provided in table 3.1 below.
[50] The reader is advised to consult Keating et al. (2015) where a far more detailed and comprehensive description of the various types of creak is provided than is warranted here.

Taking these considerations into account, I adopted a conservative approach towards the coding of breathy voice in this study. That is, if there was any doubt about whether a given segment could be classified as breathy, that segment would be coded 'modal' instead. Thus it is expected that within the modal category, there will be segments included with varying degrees of breathiness (but no segments containing any form of creak). Only those segments which were unambiguously breathy would be coded as such. This has important ramifications for the interpretation of the research findings as presented and discussed in the following two chapters.

The guideline for coding used containing all of the phonation type categories used in the auditory coding procedure is provided in table 3.2 below, which had been devised based on the relevant literature.

Code	Voice quality	Basic description
M	Modal	Regular vibration of the vocal folds with F0 neither particularly high nor low
B	Breathy	Aspiration noise with voicing (the aspiration noise should be visible in the spectrogram and waveform) (see Podesva 2013)
C	Creak (prototypical)	Low and somewhat irregular F0 accompanied by the auditory impression of creak (see description in Keating et al. 2015)
F	Vocal fry	Low but not necessarily irregular F0, with very high damping between pulses and a very clear 'picket fence' effect (see description in Keating et al. 2015)
W	Whisper	Aspiration noise without voicing (see Podesva 2013)
Cw	Creak+whisper	Creak with some accompanying aspiration
Fw	Vocal fry+whisper	Vocal fry with some accompanying aspiration
H	Shimmer and/or jitter	Irregular pitch or waveform amplitude accompanied by an impression of harshness or roughness
U	Unidentified	Any clearly non-modal voice quality not falling into any of the other categories

Table 3.2: Coding categories and basic descriptions used as a guide in the auditory coding of the sentence data.

The extracted vowel samples from these speakers were presented to the trained research assistant as well as to myself in a random order on a data disc. This helped prevent bias in coding, since many of the extracted tokens were too short in duration to be identified as belonging to a particular speaker when presented in a random order. This allowed for a measure of inter-rater reliability to be calculated for the auditory analysis procedure as will be reported in the following chapter.

3.9. COMPARISON AND STATISTICAL ANALYSIS USING LINEAR MIXED EFFECTS REGRESSION AND WILCOXON RANK SUM TESTS IN R

Following extraction of the parameter estimate values from VS, the data were copied and pasted into an Excel spreadsheet formatted in a way which would allow it to be easily transferred as a dataframe into the statistical analysis program R. Graphs were subsequently generated based on these data, which are displayed in the following chapter and which were used to illustrate general trends present in the data.

In the interview data analysis which forms the main analysis used in this study, I excluded all unstressed vowels as these were typically too short for reliable VS measurements to be taken. I likewise removed vowels shorter than 60ms for essentially the same reason, following Esling and Edmonson (2011) and Newman and Wu (2011). I also excluded segments for which VS did not obtain an RMS Energy reading, since such segments are unlikely to yield reliable measurements for the other parameter estimates. Following these exclusions, the total number of vowel tokens used for the analysis was 9332, with an average of 259 tokens per speaker for the interview data.

While I closely followed the structure of the linear mixed effects model used by Szakay and Torgersen (2015), I also adapted it somewhat to better suit my data. For example, since only female speakers were included in my study, gender was not entered as a fixed effect in my linear mixed effects model.

In order to ensure greater comparability with Szakay and Torgersen (2015) and in order to control for the influence of outliers, for the interview data, I removed data points 2.5 standard deviations away from the mean for each measure.

As was done in Szakay and Torgersen (2015), I also log transformed the values for continuous variables before entering them into the linear mixed effects model. This was to ensure comparative scale among the measures. However unlike in Szakay and Torgersen (2015) I only log transformed the independent variables and left the dependent variables untransformed.

This is because while the independent variables of fundamental frequency, first formant frequency, intensity and vowel duration (after the abovementioned exclusions had been implemented) do not have zero values, dependent variables such as H1−H2 for example (and the

other dependent variables) not only have zero values, but also have negative values and both the zero and negative values are meaningful in this case. Such variables as H1−H2 with negative values as well as zeroes are not suitable for log transformations because it is not possible to compute logarithms for zero values and negative numbers (Baayen 2008; Sheskin 2007:465).

Baayen (2008:31) points out that logarithmic transformations have the advantage of eliminating or at the very least reducing distribution skewing substantially to prevent extreme outliers from dominating the outcome or potentially completely concealing those main tendencies which characterize most of the data points. A number of functions in R, including the linear model function may also not function correctly if skewness is not dealt with by means of such transformations as the log transformation (Baayen 2008:92). The intercept and slope in the regression models may become shifted to such a degree by overly influential but high probability outliers, that for most data points, the model becomes suboptimal (Baayen 2008:92). Baayen (2008:92) therefore suggests the use of the logarithmic transformation as the technical solution to this problem such that the regression line is able to capture the main data trend.

As pointed out by Kirk (2013:106) logarithmic transformations typically have the benefit of both a reduction of the influence of more extreme scores as well as making scores resemble the normal distribution to a greater degree.

As Kirk (2013:104) notes, reducing the effect of extreme scores is useful since these are undesirable for both substantive reasons as well as for statistical reasons (namely that their presence decreases statistical power while increasing variability).

It is possible to log transform both dependent and independent variables which would entail adding an arbitrary constant value. It is also possible to log transform only the independent variables. Although it is sometimes recommended that a constant of 1 (or sometimes 0.05) be added in such cases, it has been pointed out (Berry 1987;Yamamura 1999) that the exact magnitude of any constant added in this way may have an effect on the conclusions reached, making the addition of such an arbitrary constant undesirable. For the variables to be transformed in this study, the exact value for this constant would also need to differ depending on which variable was being logarithmically transformed. I therefore selected the second option, namely to log transform only the independent variables (thus avoiding the addition of an

arbitrary constant value). The interpretation of log transformed results where only the independent variables are log transformed is relatively straightforward, namely a 1% increase in the independent variable decreases or increases (depending on the coefficient, β, for the regression) the dependent variable by a number of units equal to the coefficient divided by 100 (Introduction to SAS. UCLA: Statistical Consulting group 2018). That is, where there is a 1% increase in for example, logpF0 (the log-transformed values for fundamental frequency) this would results in a coefficient/100 increase/decrease in the dependent variable (for example, H1*−H2*.

While there were a variety of options available for the statistical analysis of the data, the method which was chosen as most appropriate was that of linear mixed models regression.

Several recent sociolinguistic studies investigating voice quality variation have utilized this type of statistical analysis, thus illustrating the suitability of this form of statistical analysis for testing the strength of different predictors of voice quality variation in sociophonetic studies generally and specifically for studies involving the investigation of the differential use of phonation types by different groups. These studies are briefly reviewed here.

For example, Podesva (2013), in his investigation of gender and the social meaning of non-modal phonation types used a mixed effects linear regression for each of the phonation types being investigated and factorially crossed the social factors of the study (namely age, race and sex). The fixed effects included in this model were the length of the intonation phrase in syllables as well as whether or not a given phrase included reported speech, while the random effect was that of the individual speaker (Podesva 2013).

Likewise, Gaither et al. (2015) also used linear mixed-effects regressions in their study of the priming identity and biracial speech. Gaither et al. (2015) claim that using this method of statistical analysis allowed them to control several important potentially confounding variables as well as to provide greater statistical power.

Most recently, Szakay and Torgersen (2015), in a study investigating the effect of ethnicity, gender and f_0 in an analysis of voice quality in London English, also made use of a

linear mixed effects model for part of the statistical analysis of their data. This mixed effects linear regression model was used for the harmonic differential (H1−H2) data for the Hackney subsample (one of their regional subsamples for which data existed for both Anglo and non-Anglo speakers) (Szakay and Torgersen 2015). In this analysis, vowel duration, F1 values, intensity, f_0, ethnicity and gender were entered as the fixed effects, while the random effects entered were 'word' and 'speaker' (Szakay and Torgersen 2015: 2-3). In order to ensure a comparable scale among continuous variables, these were entered as log values in the model (Szakay and Torgersen 2015).

In my statistical analysis of the interview data using linear mixed effects regression, I followed a similar model to that of Szakay and Torgersen (2015) in order to allow for more comparability between the results and because the research aims as far as the purpose of the use of the linear mixed effects regression analysis is concerned, are similar. For the exploratory sentence data, a different model was used since these data essentially constituted a separate data set which was useful for exploring the interaction effects between vowel quality (for this data set only three vowel phonemes were used and so vowel quality was set as one of the predictors) and the measures as well as the patterning of the measures for the auditorily identified phonation types. The model used for the interview data has a more direct bearing on the matter of ethnolinguistic differences in long-term voice quality and closely follows the model used by Szakay and Torgersen (2015).

For their linear mixed effects model, Szakay and Torgersen (2015) used H1−H2 as the dependent variable with continuous independent variables of fundamental frequency, intensity, first formant frequency as well as vowel duration. They used the categorical variables of gender, ethnicity, speaker and word (Szakay and Torgersen 2015). While following this model closely, it was necessary for me to adapt the model to suit my dataset and to ensure greater accuracy. Naturally, given that all of my research subjects were female, I did not include gender as an independent categorical variable as mentioned earlier. I used the RMS Energy measure supplied by VS in place of Szakay and Torgersen's (2015) intensity measure and entered it as a predictor variable in the model. Due to the fact that it would be helpful to account for the effect of F2 as one of the acoustic measures relevant to the potential influence of vowel quality, I included F2

along with F1 as an additional continuous predictor variable in my model. The categorical variable of 'word' was permitted to have a random intercept to model the different baseline values for the parameter estimates for each word. Thus the model expects multiple responses for each word and that these depend on the baseline for that word. This model therefore accounts for non-independencies (the model assumes multiple responses for each word and for each subject) as well as accounting for by-word and by-subject variation in the overall values for the parameter estimates (Winter 2013:5).

The categorical variable of 'speaker' was permitted to have both a random slope as well as a random intercept. As noted by Winter (2013:15), in random intercept models, it is assumed that the effect (in the case of this study the primary effect of interest being that of ethnicity) is the same for both subjects as well as words. However, this is not an entirely valid assumption because it is expected that the effect of ethnicity may well be different for different subjects, although given the diversity of words occurring in the interview and because these are likely to differ drastically in terms of overall context and phrasal position it is not likely that particular words would elicit more creaky voice or more breathy voice for example based on the ethnicity of the speaker, that is the effect of ethnicity should not be different for different words if it is indeed a voice quality effect according to how voice quality has been defined in this study. However, we might well expect that the effect of ethnicity may be greater for some speakers than others. For example, for some speakers the signaling of ethnolinguistic identity by linguistic means may be more important than for others. For this reason, while a random intercept is sufficient for the variable 'word,' speakers should have both a random intercept as well being permitted to have random slopes for the effect of ethnicity. That is, the model expects that there should be different baseline levels for each speaker as well as different responses between individual speakers for the effect of ethnicity.

The use of mixed effects models incorporating random slopes as justified by the research design has the advantage of avoiding the fairly high Type I error rates associated with the use of such models without including random slopes (see Schielzeth & Forstmeier 2009 and Barr Levy, Scheepers and Tilly 2013).

The log transformed values for pF0 (fundamental frequency estimated using the PRAAT algorithm as implemented in VS), RMS Energy, pF1 (PRAAT algorithm-derived measure of F1 in VS), pF2 (PRAAT algorithm-derived measure of F2 in VS) and duration were entered as the continuous predictor variables in order to take account of the possible influence of pitch, intensity, vowel quality and duration in the model. I also tested for interaction effects between the variable of ethnicity and the continuous predictor variables, duration, RMS Energy, f_0, F1 and F2 using the log-transformed values for these measures. In order to test for significance, both for the interactions and for the main effect of ethnicity, I employed likelihood ratio tests using ANOVAs.

The principle of the likelihood ratio test is to make a comparison of the likelihood (the probability that one would observed the data collected based on the model) of one model with that of the other (Winter 2013:11). In this case, it would be a comparison of the model without the factor of ethnicity and the full model incorporating the factor of ethnicity. By using the anova () function in R for these two models, it is possible to derive *p*-values indicating the significance of the effect of ethnicity. The same principle applies for the testing of significance of interactions. If the results are significant when comparing the full model to the interaction model by means of an anova, it is possible to conclude that for example, there is a significant inter-dependence between ethnicity and fundamental frequency for one of the parameter estimate measures (Baayen 2008; Winter:2013:14).

As Sheskin (2007:1610) notes, the likelihood ratio test can be used whenever there is a need to compare two models with one another. McCulloch and Searle (2001:24) point out that the F-statistics used in the hypothesis testing in the traditional ANOVA methodology based on assumptions of normality can be shown to be a result of the likelihood ratio test, as first proposed by Neyman and Pearson (1928), the application of which is much broader than the traditional ANOVA. Baayen (2008:253) states that likelihood ratio tests are used to assess the significance of random effects parameters as carried out by the anova () function in R. Two mixed-effects models must be provided for the anova () function and these models should be alike in terms of the structure of the fixed effects but differ with respect to the number of random-effects (Baayen 2008:253).

Thus for example, I would first build the full model and test it by means of an anova test to test for the significance of the variable between the interaction model and the full model.

Since it is suitable for variables with skewed distributions (Baayen 2008:76) and in cases where the distributional assumptions (symmetric and unimodal) of the *t*-test are violated (Everitt and Hothorn 2006:27), it has greater statistical power than most other non-parametric tests as well as to ensure some level of comparability with Szakay and Torgersen (2015), I used the Wilcoxon rank sum test as implemented in R in order to test overall differences in terms of ethnicity.

The Wilcoxon rank sum test has both the advantage of not being likely to be effected by outliers as well as not requiring the assumption of a normal distribution as would be necessary for a *t*-test (Everitt and Hothorn 2006:30; Sheskin 2007:513) due to the fact that it is based not on the observations themselves but on their joint ranks for both groups. The Wilcoxon rank sum test is one of the most sensitive tests for the testing of differences in location (i.e. central tendency) between two samples (Siegel and Castellan 1988:166). Due to its statistical power (with a power-efficiency close to 95% even for samples of moderate size), it is "an excellent alternative to the t test" (Siegel and Castellan 1988:137) without requiring all of the restrictive requirements and assumptions of the *t*-test and for certain cases may in fact have greater statistical power in comparison to the *t* test .

I conducted the Wilcoxon rank sum tests for the hypothesis that there was a difference as well as for the hypotheses that the difference was in a certain direction, both greater or less for the one ethnic group than for the other.

The Wilcoxon rank sum test as implemented in R where both and x and y (in this case, datasets for black and white speakers for a particular VS measure respectively) are provided and the option 'paired' is set to FALSE conducts an equivalent of the Mann-Whitney test. The null hypothesis is thus that the distribution for black speakers differs from that of white speakers by a location shift of μ, with the alternative hypothesis that they differ by another location shift (R Core Team 2016). One-sided alternative hypotheses were also tested for each dependent

variable. The alternative hypotheses for the one-sided tests stated that that the distributions for black speakers were shifted to the right of that of white speakers and that the distributions for black speakers were shifted to the left of those of white speakers[51]. All significant results for these tests are reported in the following chapter.

3.10. CONCLUSION

In this chapter, I have provided a detailed description of the sampling and recording procedures, as well as the acoustic, auditory and statistical data analyses used in this study in order to meet the stated research objectives. As part of this description, I have also reviewed the literature concerning the techniques and measures used in this study where appropriate. The following chapter documents the results based on the analyses described in this chapter.

[51] For more on the Wilcoxon rank sum test as implemented in R, interested readers are encouraged to consult the R documentation (R Core Team 2016).

CHAPTER IV: RESULTS FOR THE AUDITORY ANALYSIS AND THE HARMONIC DIFFERENTIAL MEASURES OF THE ACOUSTIC ANALYSIS

4.1. INTRODUCTION

In this chapter I first present the findings of the auditory analysis of the sentence data and I subsequently present the findings of the acoustic analysis of the interview data, which relates more directly to the phenomenon of voice quality and therefore forms the main focus of the current study. In this chapter, I also provide some interpretation of the research findings based on the relevant literature. In the acoustic analysis section, I present the findings for the measures of spectral balance first, followed by other spectral amplitude measures. In the following chapter, I provide the results of the acoustic analysis for the noise measures.

In the acoustic analysis section, for each measure, I provide a graphical display of how the measure relates to the phonation types used in the auditory analysis of the sentence data. I subsequently present the findings of the statistical analysis for both the interview and sentence data, with a greater focus on the results for the interview data. In cases where there are important interactions between some of the variables and ethnicity, I have also included scatterplots to further explore some of the more interesting interactions, where the correlations in question are relatively strong. I also provide a discussion of the interactions in question. I present scatterplots of weaker, but nevertheless interesting correlations in the appendix. I have provided these since they may prove useful to future researchers. Due to limitations of space, I have not included tables of basic descriptive statistics for each measure in this chapter. I have however included these in appendix C for the sake of accountability. In the following section, I present the findings of the auditory analysis of the sentence data, which is meant to provide an overall picture of the proportional use of the auditorily identified phonation types for the speakers included in my sample.

4.2. AUDITORY ANALYSIS OF THE SENTENCE DATA

As described in the previous chapter, an auditory analysis was performed for the sentence data. This analysis is intended as an aid in identifying overall patterns in the data and as a means of potentially linking any observed acoustic differences, which form the main focus of this study, to perceptually relevant differences in voice quality within the context of South African English. Thus while the auditory analysis results do not form a definitive part of this study, they are useful in providing an overview regarding potential differences in terms of voice quality present in the data.

To briefly summarize the approach used for the auditory analysis component of the research here for the benefit of the reader, each measurement point of the read sentence data was coded as to whether during that measurement point, a particular phonation type (out of those described in the previous chapter) could be auditorily detected. The following table summarizes the number of measurement points coded according to each of the auditorily identified phonation types for black and white speakers respectively for all of the sentence data.

Phonation type / Ethnicity	Breathy Voice	Creaky Voice	Whispery Creak	Vocal Fry	Whispery Vocal Fry	Harsh Voice	Modal Voice	Whisper
white	2,730 (0.006%)	25,295 (0.057%)	214 (0.0005%)	9,829 (0.022%)	0 (0%)	5,654 (0.013%)	401,263 (90.11%)	341 (0.0008%)
black	9,267 (0.018%)	28,999 (0.057%)	8,037 (0.016%)	1,895 (0.004%)	133 (0.0003%)	3,236 (0.006%)	458,138 (89.56%)	1,810 (0.004%)

Table 4.1: The number of data measurement points coded according to auditorily identified phonation type for each of the ethnic groups for all of the read sentence data.

The data displayed in the table exhibit some interesting patterns which call for further comment. Firstly, it is immediately apparent that the overwhelming majority of measurement points for both ethnic groups are coded as modal (M). This similarity is not entirely unexpected, given that overall, the perceptual difference between the two groups in terms of phonation seemed to be minimal and that few authors had claimed that there were very obvious differences in terms of phonation as stated in previous chapters. It is worth noting that while the frequency of

use of modal voice is more or less similar across the two ethnic groups, black speakers appear to make use of modal voice slightly less frequently than their white counterparts do. It is important to once again note that the coding procedure employed is likely to have underestimated rather than overestimated the extent of breathiness in the sample, given the difficulty in confidently coding measurement points containing greater or lesser degrees of aspiration as unambiguously breathy. Thus the modal category may contain measurement points for which some aspiration noise is present.

Non-modal phonation types display differing patterns of distribution for the two groups. Whisper (W), for example, appears to be seldom used by either ethnic group, although within the category of whisper, black speakers are found to make proportionally greater use of this phonation type. It is also evident that black speakers make more frequent use of breathy voice (B) than white speakers do, by a relatively large margin, although when compared to the similarity in terms of the use of modal voice between the two groups of speakers, this difference appears relatively small. Prototypical creak (C) is used with proportionally the same frequency for both groups of speakers.

For compound creak types, such as whispery creak (CW) and whispery vocal fry (FW), frequency of use is greater among black speakers than among white speakers and whispery fry seems to be exclusively used by black speakers although this is generally quite an uncommon phonation type.

White speakers appear to use vocal fry (F) and harsh/aperiodic voice (H) with greater frequency than black speakers overall. The following figure, figure 4.1, graphically illustrates the data contained in table 4.1 above, thus displaying the proportions of the auditorily coded phonation types for the sample as a whole.

Figure 4.2 below, graphically illustrates the proportions of the auditorily coded non-modal phonation types for the sample as a whole.

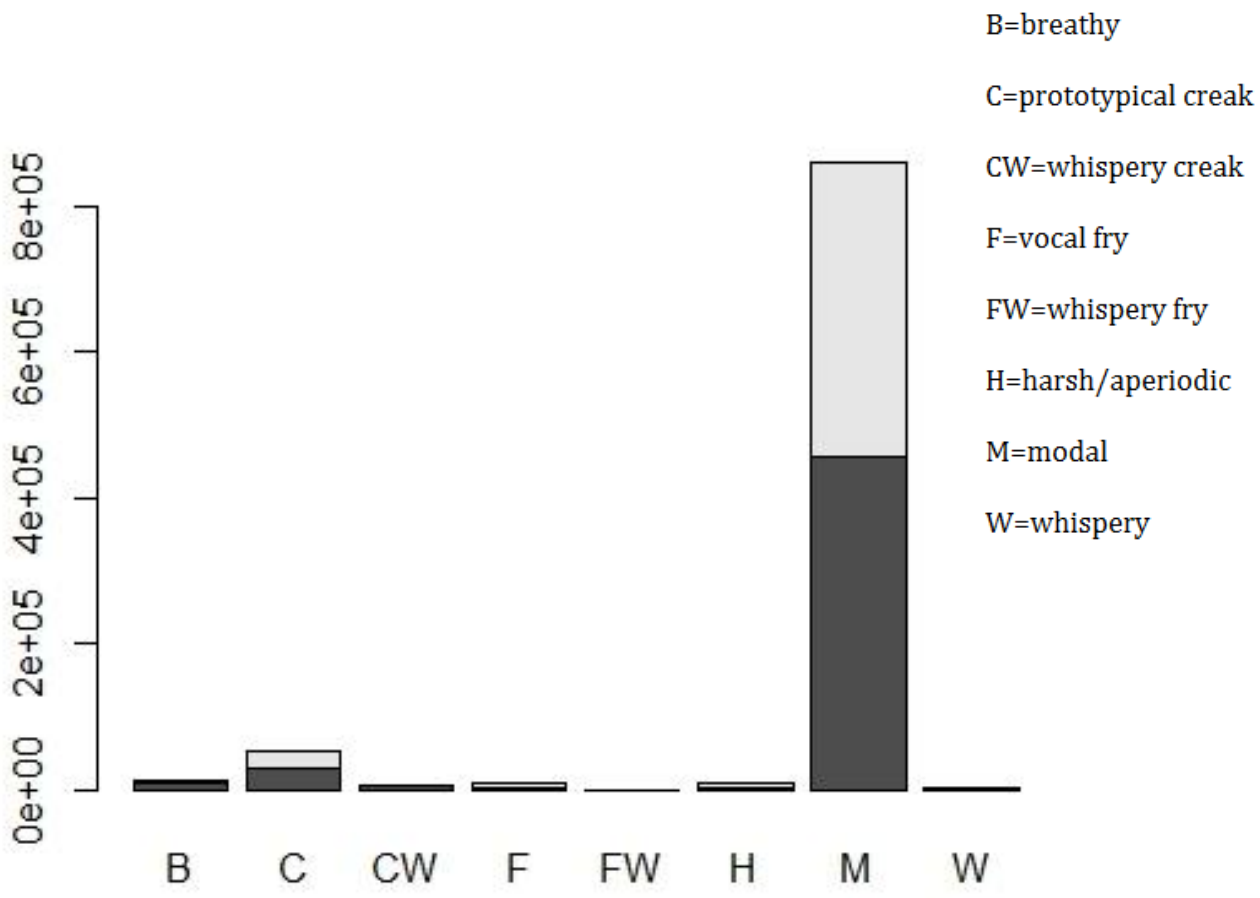

Figure 4.1: The graphic representation of the relative proportions of auditorily identified phonation types occurring in the read sentence data divided according to ethnicity (white colouring symbolizes the number of measurement points for each phonation type for white speakers and dark grey colouring the number of data measurement points for each of the phonation types for black speakers).

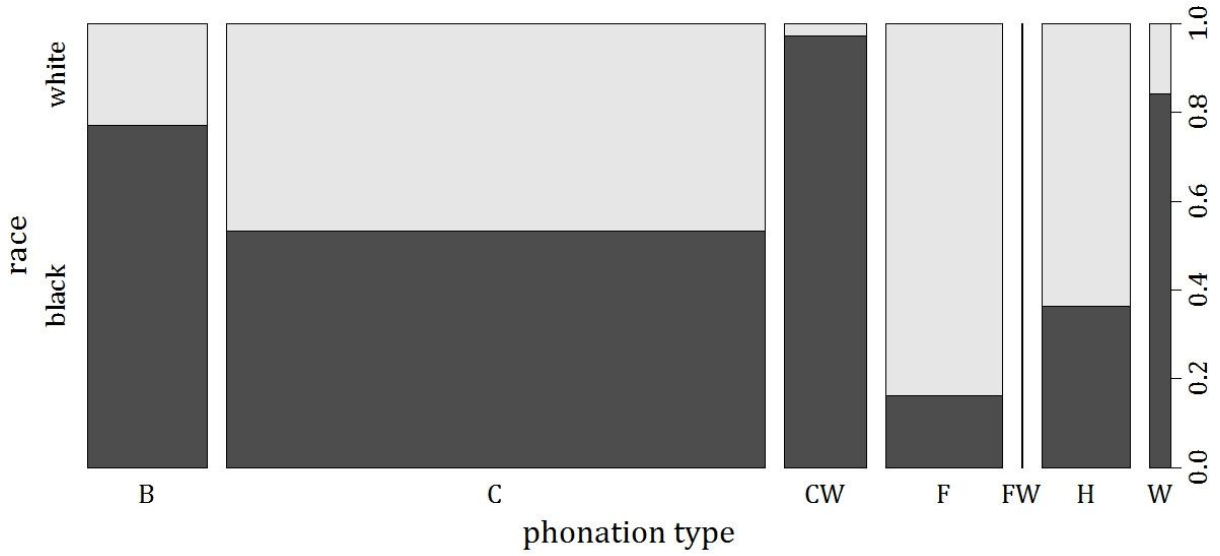

Figure 4.2: The graphic representation of the relative proportions of auditorily identified nonmodal phonation types occurring in the read sentence data according to ethnicity. B=breathy, C= prototypical creak, CW=whispery creak, F=vocal fry, FW=whispery fry, H=harsh/aperiodic voice, W= whispery.

4.2.1. Interrater Reliability for the Auditory Coding Procedure

In this section, I provide an assessment and discussion of the interrater reliability for the auditory coding of phonation types used in this study. The auditory coding procedure is described in full in the previous chapter. As stated in that chapter, while I as the principal researcher provided an auditory coding of the entire sample of speech segments, I considered it helpful to provide a measure of the reliability of the coding procedure itself, that is whether the phonation type categories could potentially be useful in future research, given that the criteria used to define some of these categories are relatively new and have been adapted for the South African English data. While the findings of the instrumental and statistical data analysis presented in the bulk of this chapter do not depend on the reliability of the auditory coding procedure as such, the impressionistic data provided by means of this procedure may supply an additional perspective which may be used in assessing and interpreting the results of the instrumental analysis and therefore some measure of the reliability of this procedure is appropriate.

To this end, as described in the previous chapter, a research assistant, Yolandi Klein, was trained in the procedures of how to code the data and provided an auditory coding of the data for four speakers, two black and two white. The measure of interrater agreement I chose is Krippendorff's alpha (Krippendorff, 2012), as computed by utilizing the 'irr' package in R. In this section I therefore report on and provide a brief discussion of the values for Krippendorff's alpha for the coded data for four speakers as coded by the principal researcher and the research assistant.

The total number of tokens coded by both raters was 862. The value of Krippendorff's alpha for these data is 0.461. This is not a particularly high level of inter-rater agreement, but is reasonably high considering various factors. The coding procedure as described in the previous chapter involved coding a given segment for more than one phonation type if more than one phonation type was present. Thus for certain segments, due to the complexity of the segment in terms of phonation, there would be more than one symbol used. This would naturally lead to lower interrater agreement than would have been the case if each segment was coded for only the dominant phonation type in that segment.

Once the number of possible symbols is reduced for each segment by reducing the number of phonation types by using broader distinctions between phonation types, there is an associated increase in interrater reliability as expected and discussed below.

It should also be noted that as described in the previous chapter, both auditory coders (the principal investigator and research assistant) were initially inexperienced in the use of this auditory coding procedure to some extent and thus the alpha values reported in this section reflect the reliability of this procedure as applied using coders who are relatively new to the practice of coding for phonatory differences. Given that the data obtained by means of the auditory coding procedure is for exploratory use as part of the research finding as well as the aforementioned factors mitigating against a higher value for interrater reliability, the Krippendorff's alpha value of 0.461 is reasonably high.

In addition, as described in the previous chapter, I also obtained Krippendorff's alpha values for regularized coding data to ascertain the extent to which there is an improvement in

inter-rater agreement when broader phonation categories are used as opposed to the finer distinctions, for example, between different types of creak.

The first regularized coding data set I tested was that for which the two primary phonation types hypothesized to involve aspiration noise (namely whisper and breathy voice) were collapsed into one category. The motivation for potentially collapsing these categories is that in some cases it is difficult to ascertain the exact moment at which voicing either begins or ends. Thus while the more experienced coder would detect the presence of three types, for example, MBW (modal voice followed by breathy voice followed by whisper), a less experienced coder may detect only two (for example, modal voice followed by either whisper or breathy voice). The value for Krippendorff's alpha which I obtained for this regularized data set is 0.506.

A second regularization was performed where the two categories of breathy voice and whisper remained as separate types, but the two non-composite creak types, namely prototypical creak and vocal fry were collapsed into one phonation type. This assessment was made based on the possibility that in some cases, particularly for relatively inexperienced raters, the distinction between these two types relies on an assessment of the regularity of pulses, thus potentially introducing a greater level of subjectivity into the coding procedure and thus reducing interrater reliability. The value for Krippendorff's alpha I obtained for this regularized data set is 0.475. When compared to the alpha value for the original data set and when compared with the value for the set regularized for aspiration types, it is clear that interrater reliability increases more for the latter than for the data set regularized for creak types. This would suggest that there is greater difficulty in reliably drawing finer distinctions between phonation types involving aspiration (such as breathy voice and whisper) than distinctions between vocal fry and prototypical creak, as collapsing these creak types did not greatly affect the alpha value.

I conducted another test using a different set of regularized data, namely where all instances of creak (thus including not only prototypical creak and vocal fry, but also harsh/aperiodic voice, as well as the composite creak types, whispery fry and whispery creak) were recoded as prototypical creak. For this regularized data set, the value for Krippendorff's alpha I obtained was 0.59. This value is higher than for the set for which the two aspiration types

were regularized, suggesting that the distinctions between the various creak types including harsh voice may be too fine-grained to be reliable when involving less experienced raters.

The final set of regularized data I tested was a combination of the regularized data sets discussed above. That is, whisper and breathy voice were collapsed into one category and all creak types were likewise collapsed into one category and this data set was then tested for reliability. Krippendorff's alpha for this data set is 0.638, which is reasonably high, attesting to the potential reliability of the broader phonation types when involving less experienced raters.

While some of the differences observed between black and white speakers for the sentence data may seem quite important, it is also necessary to consider that overall, modal voice is still the most common phonation type for both groups by quite a large margin and that there are differences in vowel duration between the two groups of speakers. Black speakers exhibit a higher number of measurement points[52] than white speakers, indicating greater vowel length overall. Data relating to differences in vowel duration are displayed in table 4.2 below, which shows data for minimum duration, the first quartile, median, mean, third quartile and maximum duration according to both ethnic group and vowel category.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Ethnicity and vowel	Number of tokens
30	112	148	174.7	203	705	black FLEECE	1491
29	96	129	144.2	181	376	white FLEECE	1486
36	129	178	189.1	235	589	black THOUGHT	1308
29	115	156	163.7	207	373	white THOUGHT	1314
27	78	99	99.29	120	204	black STRUT	1097
27	66	80	81.79	97	149	white STRUT	1077

Table 4.2: Descriptive statistics for vowel duration for the sentence data between black and white speakers for different vowels (N=36;black=18, white=18)[53].

[52] VS was set to take measurements at 1ms intervals for each token, such that the number of measurement observations is indicative of vowel duration.
[53] Note that for this exploratory dataset, we would not expect large differences in vowel quality for speakers from ex-model C schools for these lexical sets.

4.3. ACOUSTIC DATA ANALYSIS

In this section, I present the findings of the acoustic analysis of the interview data which forms the main focus of the current investigation. I deal with each of the main VoiceSauce (VS henceforth) measures in turn, first presenting the findings for the measure in question as they relate to the auditorily identified phonation types, followed by the findings for the interview data analysis which relates most directly to the question of voice quality. Statistical results from the exploratory acoustic analysis of the sentence data are also presented following the presentation of the statistical results for the interview data for each measure, for comparative purposes.

4.3.1. Measures of spectral balance

4.3.1.1. 2K−5K (the amplitude of the harmonic nearest to 5000 Hz subtracted from the harmonic nearest 2000 Hz, the latter of which is corrected for the effect of formants and their bandwidths)*

4.3.1.1.1. 2K*−5K Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 4.3 displays the distribution for the corrected VS measure 2K*−5K according to auditorily identified phonation type for the sentence data for the entire sample. It is evident that this particular measure is relatively effective in distinguishing between the different auditorily identified phonation types. It appears that the noise component introduced by phonation types involving creak has a greater impact in raising the values for this measure than do phonation types involving more aspiration noise, such as breathy voice and whisper. This is in complete contrast with what would be expected based on most of the literature and it is not entirely clear yet why the auditory data patterns in this way.

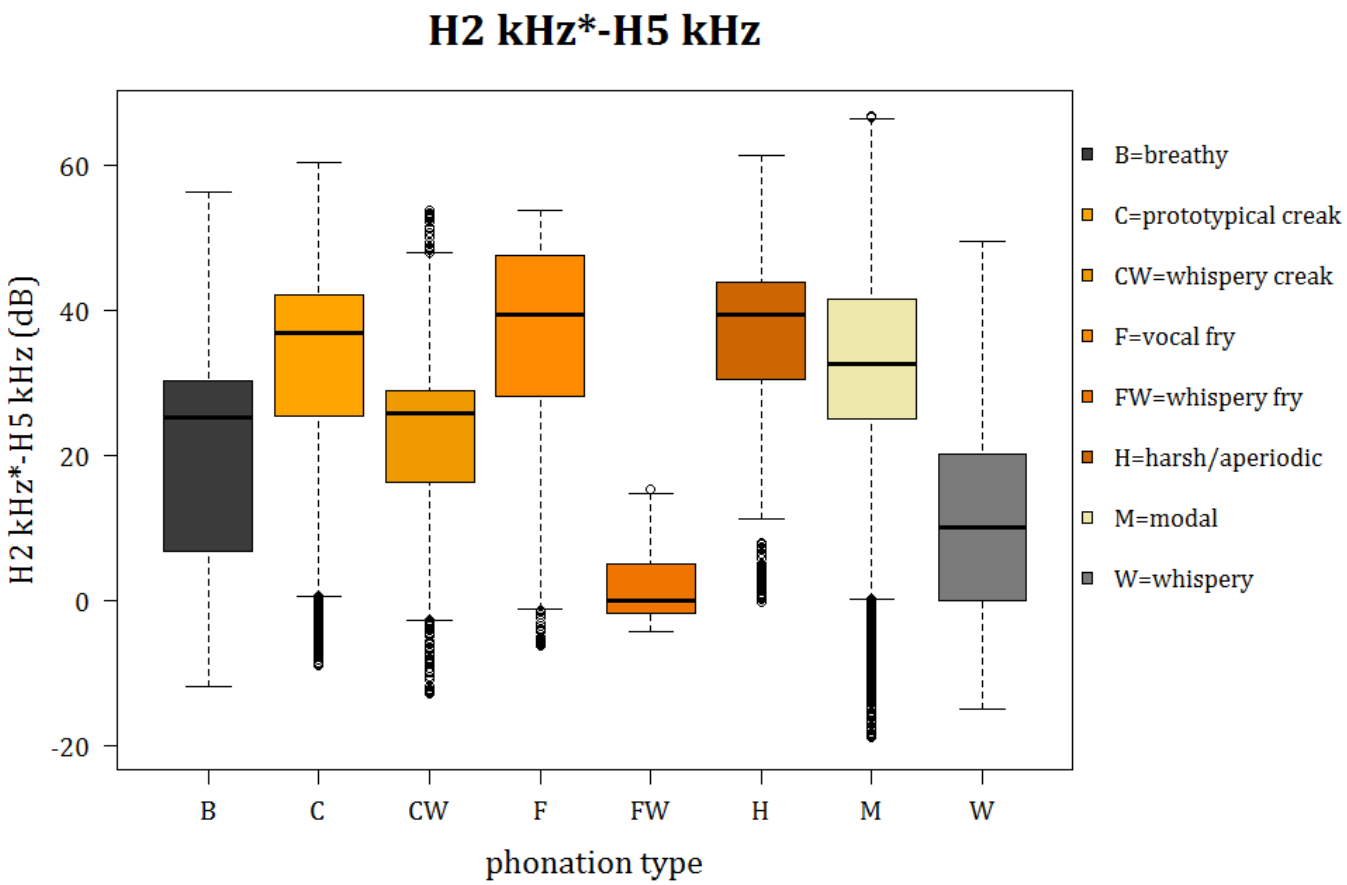

Figure 4.3: Boxplots of the distributions for the 2K*−5K data for each of the auditorily identified phonation types for all data measurement points.

4.3.1.1.2. 2K*−5K Statistical Analysis

I used R and *lme4* (Bates, Maechler and Boker 2015) to perform a linear mixed effects analysis of the relationship between the variable 2K*−5K and speaker ethnicity. This method of analysis and the reason for its selection were discussed in the previous chapter.

4.3.1.1.2.1. Interview Data

For the analysis of the interview data, as fixed effects, I entered race, logpF0 (the logged values of pF0, a measure of fundamental frequency using Praat's algorithm), logEnergy (the logged values of the root mean square energy measure), logpF1 (the logged values of pF1, a measure of first formant frequency using Praat's algorithm), logpF2 (the logged values of pF2, a measure of second formant frequency using Praat's algorithm) and logduration (logged values for vowel duration). As random effects, I entered random intercepts for speaker and for word and a random

slope for speaker for the effect of race. P-values were obtained by means of a likelihood ratio test of the full model with the effect in question against the model without the effect in question.

For this measure and in presenting the findings for all of the other VS measures included in this study, the results of the non-parametric test, namely the Wilcoxon rank sum test are presented first, followed by the results of the linear mixed effects analysis. For 2K*−5K, when testing the alternative hypothesis that the values for black speakers differ by a location shift to the right of that for white speakers (that is, that black speakers display higher values in comparison to white speakers) using the Wilcoxon rank sum tests, the difference is significant (W=178, p=0.314). This difference is also visibly evident in figure 4.4 below, which displays the values of black and white speakers for this measure for the interview data.

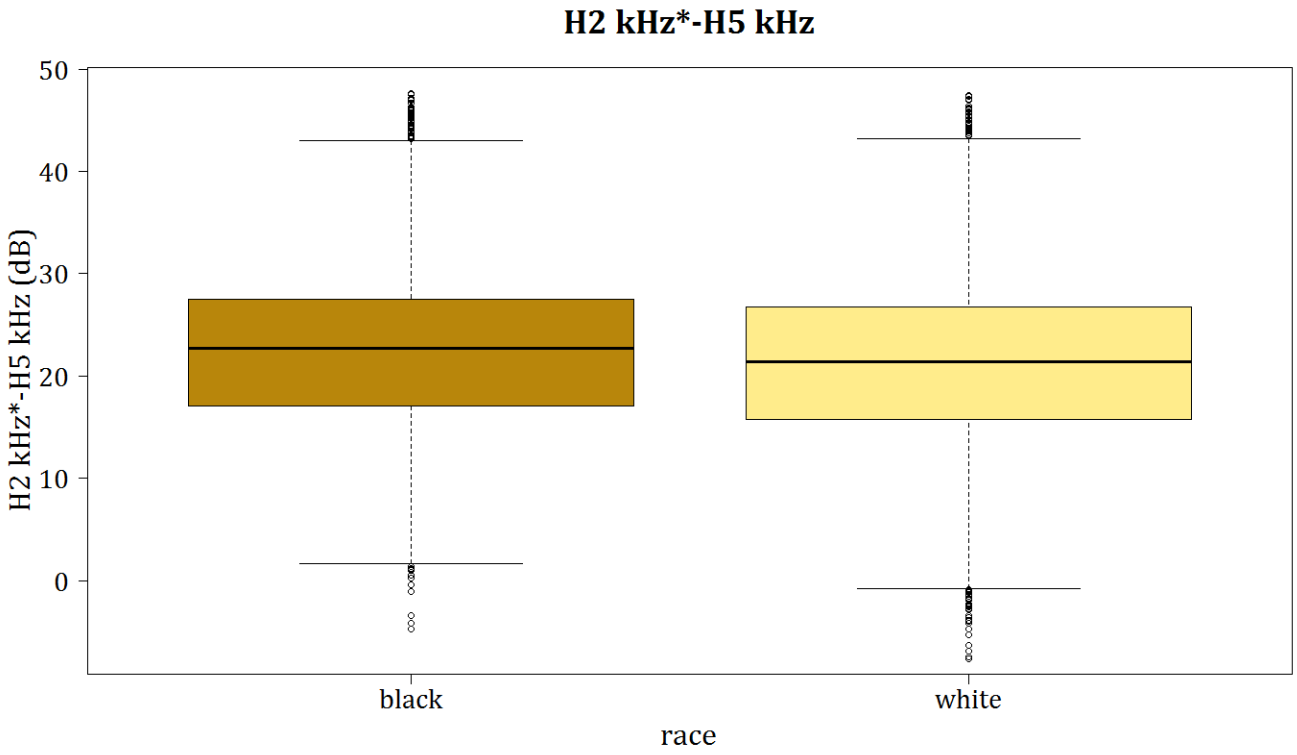

Figure 4.4: Boxplots representing the values of black and white speakers for the 2K*−5K interview data.

This distribution pattern is somewhat different to the read sentence data for the same measure as displayed below, where there are minimal differences between the two ethnic groups, although the range of values is slightly more restricted for white speakers than that for black speakers and the mean and median values are higher for white speakers. This apparent difference

(although not significant for the sentence data as pointed out, below) may suggest differences between more formal speech and more casual speech for this measure.

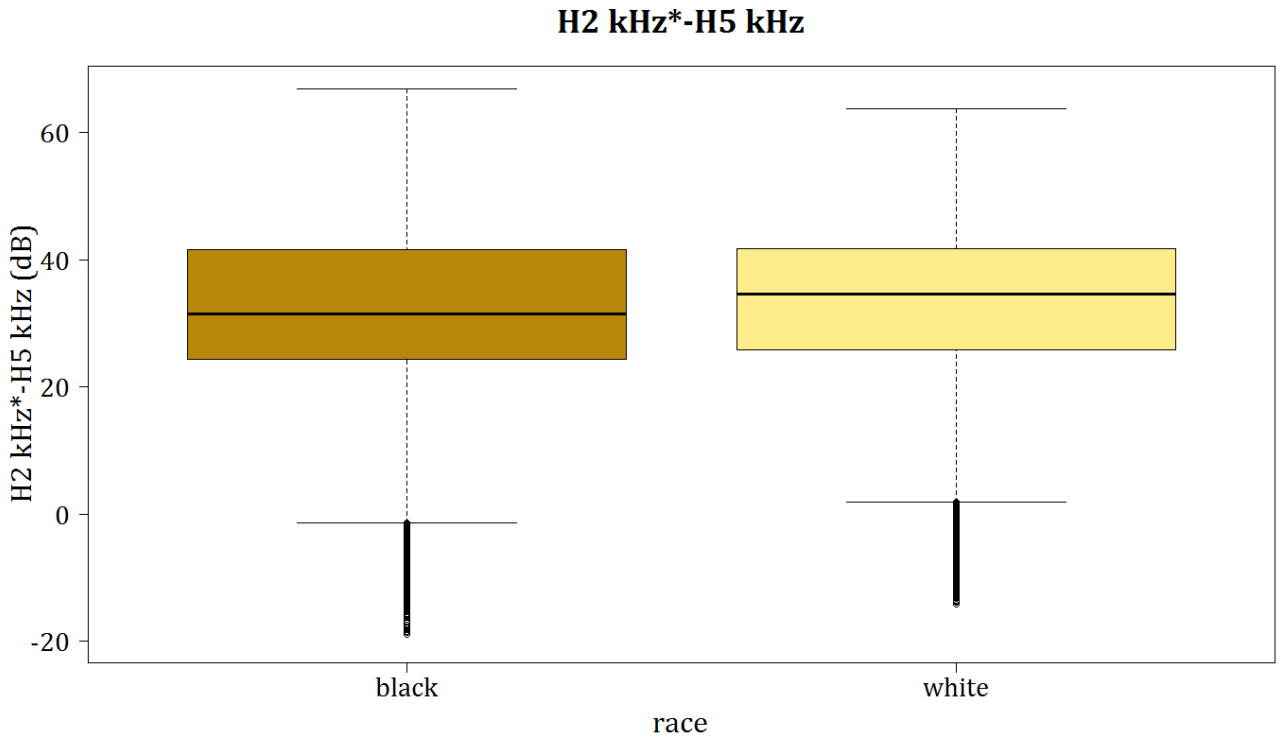

Figure 4.5: Values for the 2K*−5K read sentence data according to ethnicity.

There is a significant effect for ethnicity based on the linear mixed effects analysis results ($X^2(1)=5.543$; $p=0.019$) for the interview data, with a decrease of 1.23 dB ±0.50261 (standard errors) in 2K*−5K for white speakers.

There are also significant effects for all of the other predictors. The effect for logpF0 is highly significant ($X^2(1)= 12.818$; $p<0.001$), with a 1% increase in pF0 decreasing 2K*−5K by 0.006 dB ±0.16623 (standard errors). The effect for logEnergy is also significant ($X^2(1)= 87.838$; $p<0.001$) with a 1% in RMS Energy decreasing 2K*−5K by 0.008 dB ±0.07996 and so too are the effects of logpF1 ($X^2(1)= 3223.7$; $p<0.001$) with a 1% increase in pF1 decreasing 2K*−5K by 0.16 dB ±0.24969 (standard errors), logpF2 ($X^2(1)= 2168.2$; $p<0.001$) with a 1% increase in pF2 decreasing 2K*−5K by 0.141 dB ±0.27751 (standard errors) , logduration ($X^2(1)=79.572$; $p<0.001$) with a 1% increase in duration increasing 2K*−5K by 0.014 dB ±0.15607 (standard errors) and for speaker ($X^2(3)=517.32$; $p<0.001$).

The results of the linear mixed effects analysis of the interview data for interactions reveal a highly significant interaction between logpF0 and ethnicity for this measure ($X^2(1)= 13.455$; $p<0.001$), a significant interaction between ethnicity and logEnergy ($X^2(1)= 10.66$; $p=0.001$) and a significant interaction between ethnicity and duration ($X^2(1)= 14.15$; $p<0.001$).

The interaction between ethnicity and logpF2 approaches significance ($X^2(1)= 3.3308$; $p=0.07$), while there is no significant interaction between ethnicity and logpF1 ($X^2(1)= 0.0053$; $p=0.942$).

4.3.1.1.2.2. Sentence Data Linear Mixed Effects Analysis Results

For the exploratory analysis of the sentence data, as fixed effects, I entered race, vowel, preceding consonantal context and following consonantal context into the model. As random effects, I entered random intercepts for speakers and vowel length (denoted by seg_End in the R code), as well as by-speaker and by-vowel length random slopes for the effect of race. P-values were obtained by means of a likelihood ratio test of the full model with the effect in question against the model without the effect in question. Since, as noted in the previous chapter (and motivated therein), different models were used for the interview data and sentence data respectively, the statistical results for these two data sets cannot be directly compared. However, the sentence data results nevertheless serve as an interesting complement to the interview data which forms the main data set of interest, and so are briefly presented after the discussion of the interview data results for each measure in turn. This procedure for the sentence data applies to the linear mixed effects analysis of all of the measures included in this study subsequently discussed and therefore this information regarding how the analysis was performed will not be repeated as it applies to all of the measures discussed throughout this chapter.

For the sentence data, I found no significant effect for ethnicity ($p=0.266$), although there is a highly significant effect for the interaction between ethnicity and vowel category ($X^2(2)=7534.1$, $p<0.001$) for this dataset.

4.3.1.1.2.3. Discussion

In comparing the scatterplots illustrated in the figures in appendix E where values of 2K*−5K are plotted against pF0, for white speakers it is apparent that there is, as would be predicted based on Garellek, Samlan, Kreiman, Gerratt (2013:5), no very clear correlation between this measure and fundamental frequency, although there is a very slight positive correlation.

However, for black speakers, there is a negative correlation between fundamental frequency and this measure and the distribution appears more clearly bimodal than for white speakers when fundamental frequency is plotted against this measure.

As can be seen from figure 4.6 below, there appears to be a relatively weak negative correlation between energy and this measure overall, suggesting that louder vowels contain a weaker noise component. This would be expected, given that softer vowels may often be breathy and vowels at the end of utterances may be softer and may also contain final creak, both of which would contribute towards a greater noise component.

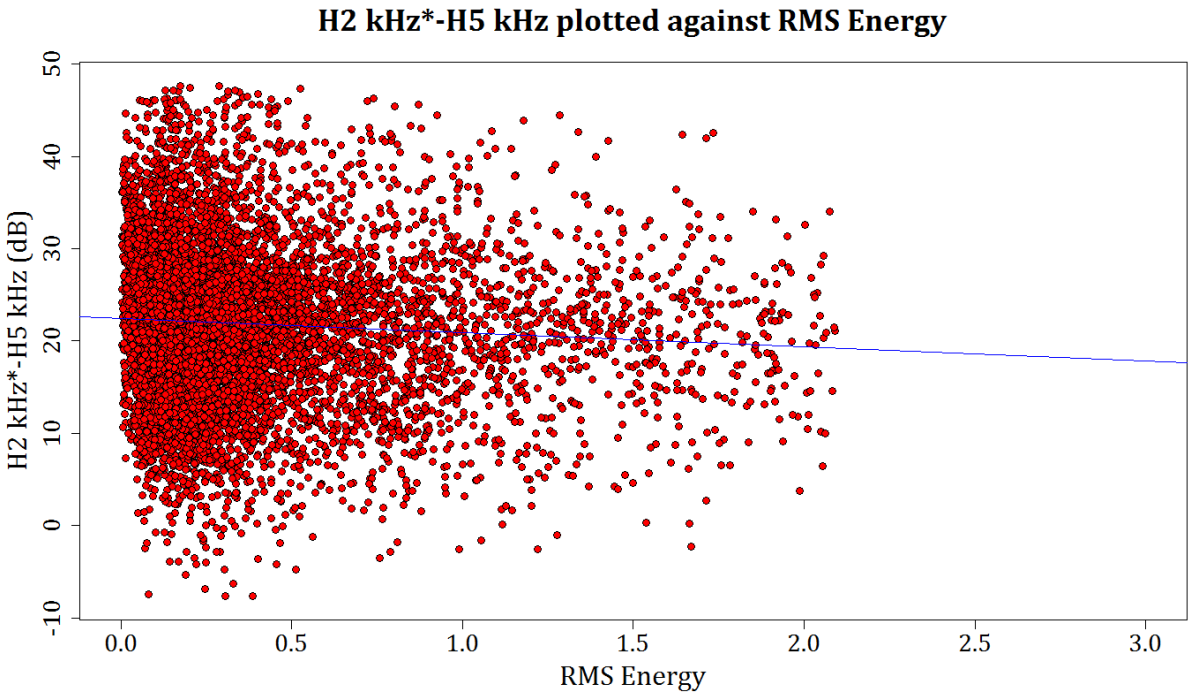

Figure 4.6: Scatterplot of the 2K*−5K data for the whole sample plotted against RMS (root mean square) energy (Pearson's $r = -0.069$).

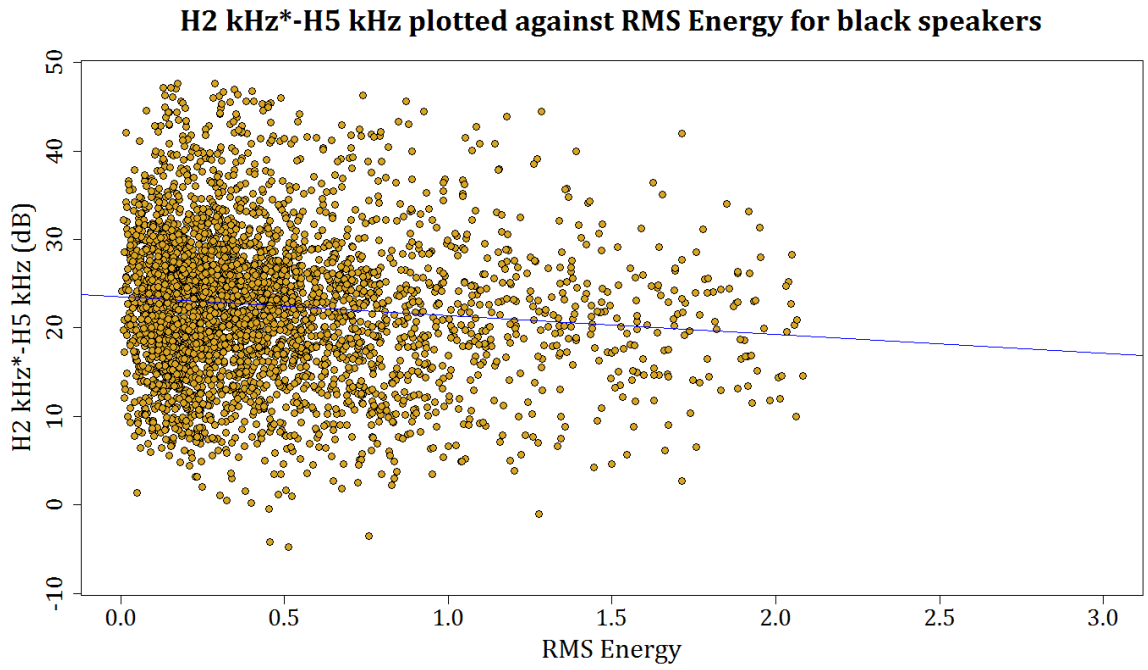

Figure 4.7: Scatterplot of the 2K*−5K data for black speakers plotted against RMS energy (Pearson's r= −0.103).

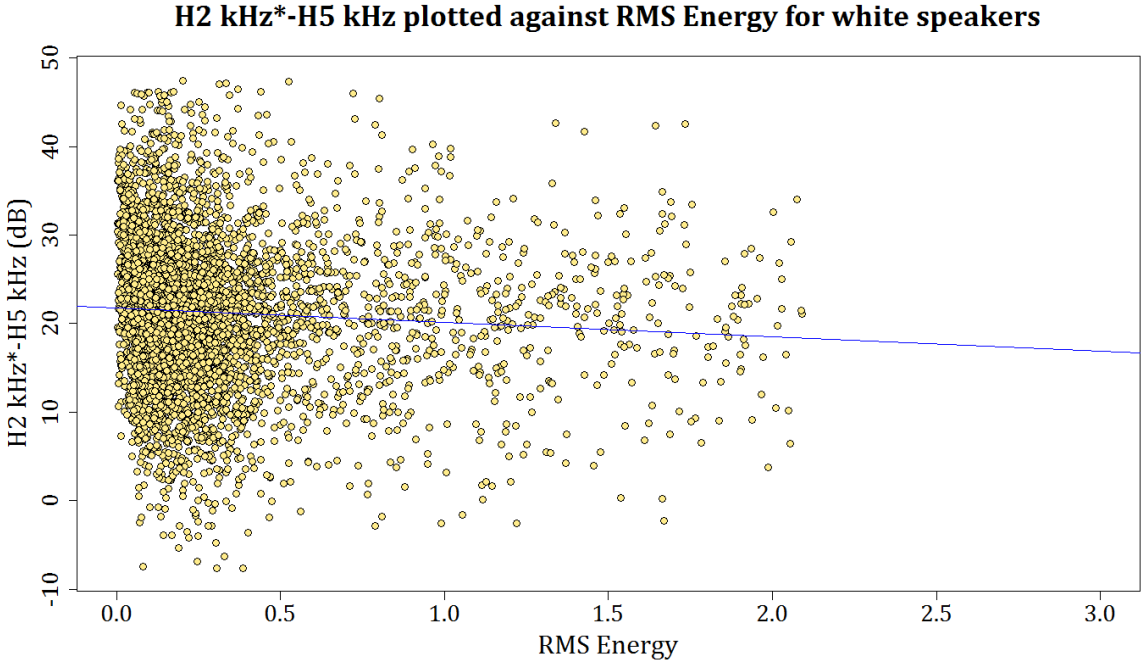

Figure 4.8: Scatterplot of the 2K*−5K data for white speakers plotted against RMS energy (Pearson's r= −0.069).

Gerratt, Kreiman and Garellek (2016) found that the mean value for H2−5kHz was slightly lower for continuous vowels (generally shorter) in comparison to sustained vowels (generally longer). That is, for longer vowels, there may be a slight increase in the predominance of the noise component. This same relationship can be observed in figure 4.9 below. This could plausibly be linked to the increased use of non-modal phonation types for longer vowels. It is also clear that the relationship is slightly stronger for black speakers than for white speakers, as can be seen from figures 4.10 and 4.11 below. The anticipated positive correlation between this measure and duration is stronger for black speakers than for white speakers.

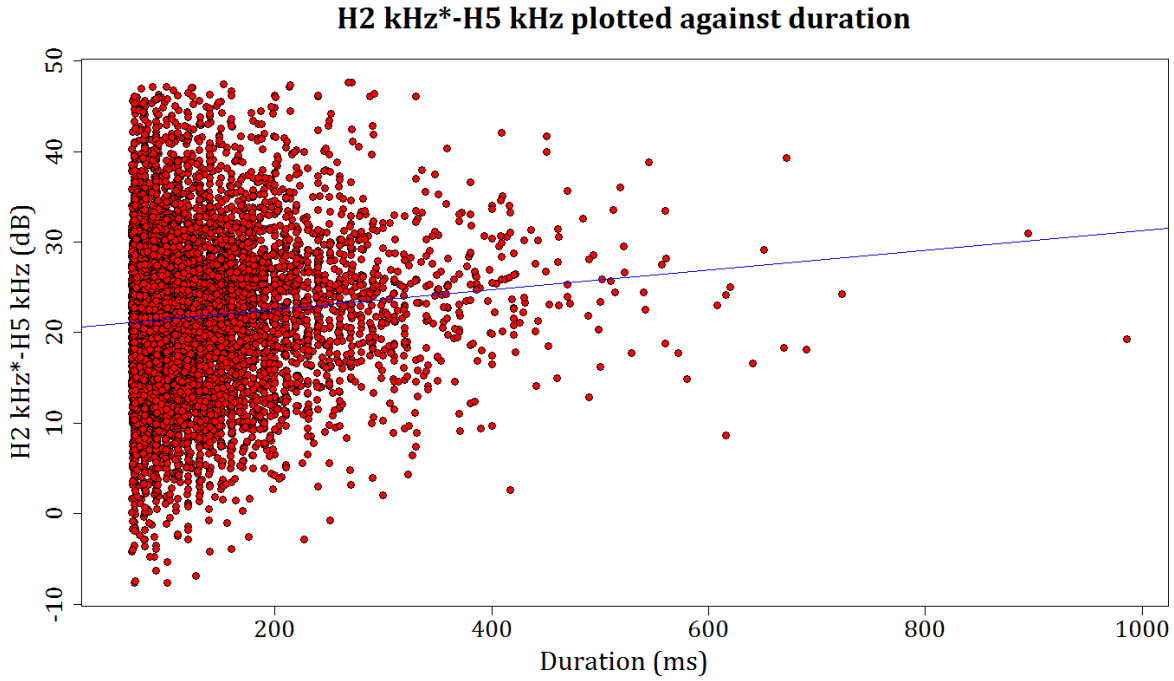

Figure 4.9: Scatterplot of the 2K*−5K data for the whole sample plotted against duration in milliseconds (Pearson's *r*= 0.093).

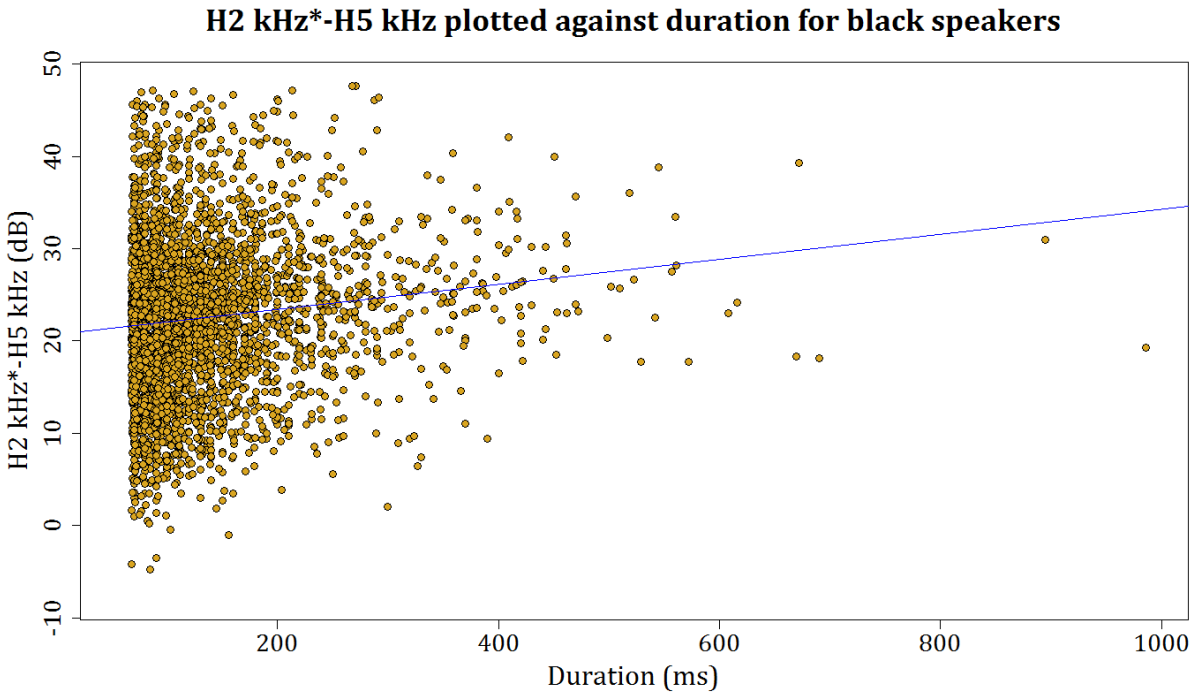

Figure 4.10: Scatterplot of the 2K*−5K data for black speakers plotted against duration in milliseconds (Pearson's *r*= 0.127).

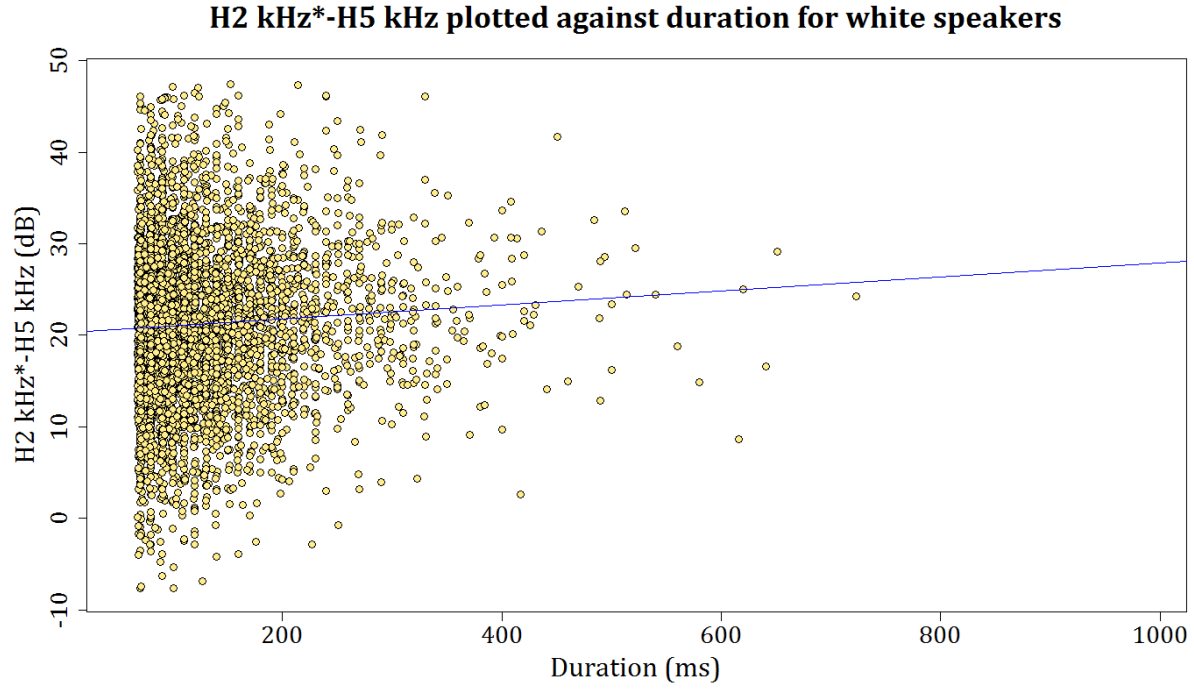

Figure 4.11: Scatterplot of the 2K*−5K data for white speakers plotted against duration in milliseconds (Pearson's *r*= 0.06).

As can be seen from the scatterplots presented in appendix E, there are also interactions in evidence when 2K*−5K is plotted against the values for second formant frequency. The pattern reveals an abrupt change in slope around 2000 Hz (as described by Kreiman et al. 2011 and Kreiman and Gerratt 2011), and so it is not amenable to representation by means of a linear correlation between 2K*−5K and F2.

4.3.1.1.3. Summary of the Findings for the 2K*−5K Measure

The Wilcoxon rank sum test results for this measure reveal a significant difference between black and white speakers, such that the values for black speakers are higher overall in comparison to those for the white speakers. The linear mixed effects analysis also reveals a significant effect for ethnicity for this measure, with the effect of white ethnicity being to decrease 2K*−5K values. The effects for all other predictors are also statistically significant for this measure. There are a number of significant interactions between ethnicity and the other predictors, including fundamental frequency, energy and duration, with the interaction between ethnicity and second formant frequency approaching significance. There are some interesting patterns with regards to some of the interactions. While these results are in agreement with the overall pattern for black speakers for the sentence data in table 4.1, they are in conflict with the results of the auditory analysis, as presented in figure 4.3, which shows that lower values for this measure are associated with auditorily identified phonation types characterized by more aspiration noise. This finding (as illustrated in figure 4.3) is unexpected and its source is unclear and perhaps worthy of further research in future.

Based on the established correlations with this measure however, the overall findings suggest that there is a greater noise component for black speakers than for white speakers, although, given the distribution pattern for the auditory analysis data for this measure, this difference could arise from either more extensive use of creaky phonation types by black speakers or greater use of breathiness or some combination of these.

4.3.1.1.4. Summary of Results for 2K−5K (the uncorrected equivalent measure of 2K*−5K)

While a significant effect for ethnicity was found for the Wilcoxon rank sum tests, revealing that the values are higher for black speakers overall, there is no significant effect for ethnicity according to the linear mixed effects analysis, which can presumably be attributed to the fact that significant effects were found for all the other predictors and the effects for these predictors are

presumably strong enough to outweigh the importance of ethnicity as a predictor for this measure. I also found significant interactions between ethnicity and both F1 and duration for 2K−5K.

4.3.1.2. H4−2K* (the fourth harmonic minus the amplitude of the harmonic nearest 2000 Hz, both corrected for the influence of formants and their bandwidths)*

4.3.1.2.1. H4*−2K* Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 4.12 displays boxplots depicting the values for H4*−2K* data according to auditorily identified phonation type for all data measurement points for the sentence data.

All phonation types hypothesized to involve a higher degree of aspiration noise display relatively high values for this measure. These include whispery fry (which has the highest values), breathy voice, whispery creak and whisper, which all have somewhat similar values for this measure in terms of displaying in general, higher values. Auditorily identified phonation types which are hypothesized to involve aspiration noise to a lesser extent, such as modal voice, prototypical creak, vocal fry and harsh/aperiodic voice have lower values for H4*−2K* in general and all have similar values in terms of standard deviation, although the median for modal voice is higher than for the other phonation types.

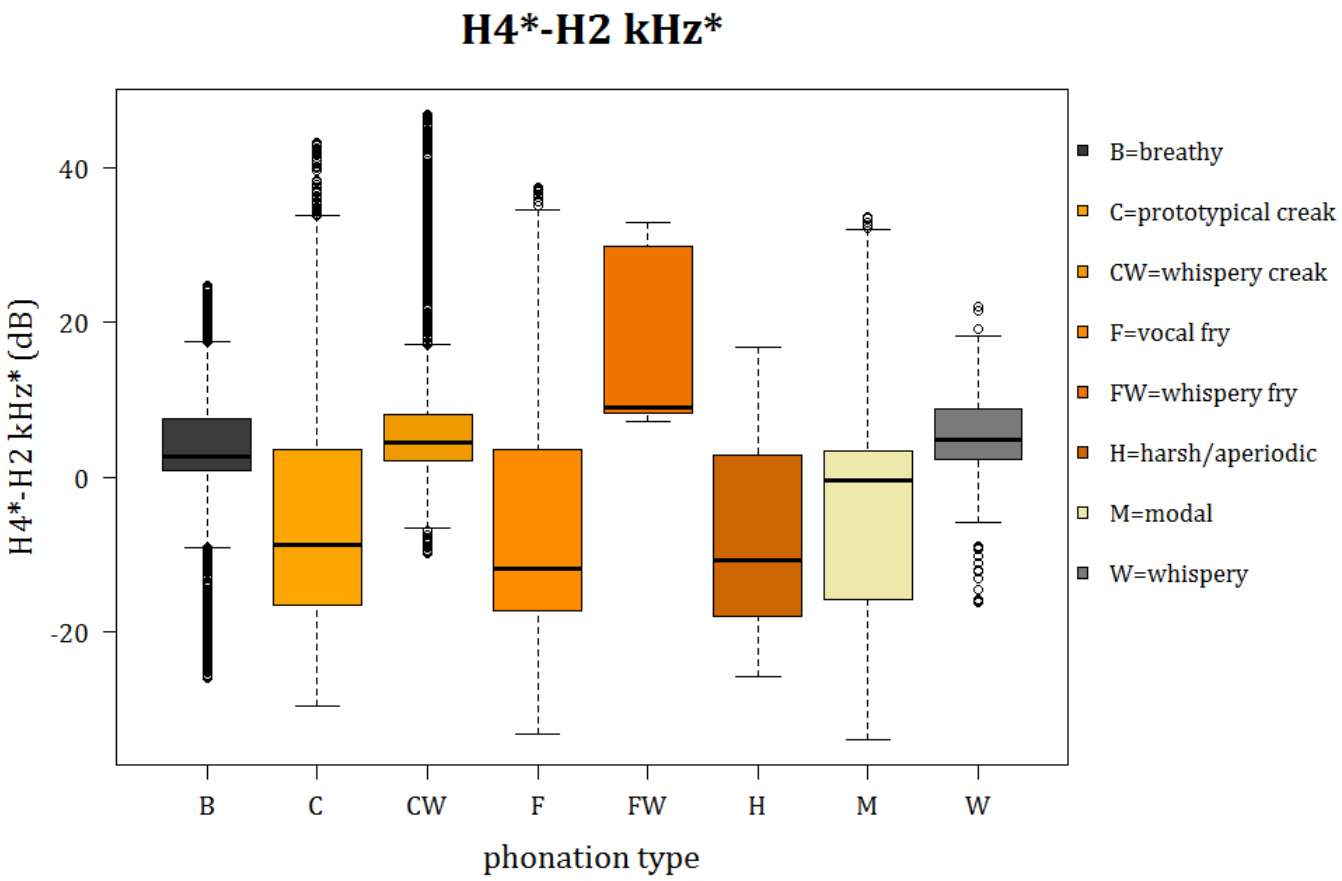

Figure 4.12: Boxplots displaying the values for the H4*−2K* sentence data for each of the auditorily identified phonation types for all data measurement points.

4.3.1.2.2. H4*−2K* Statistical Analysis

4.3.1.2.2.1. Interview Data

There are significant results for the Wilcoxon rank sum test for the interview data, for the alternative hypothesis that white speakers have higher values (W=95, p=0.011), although this is not very clear from the boxplot displayed below.

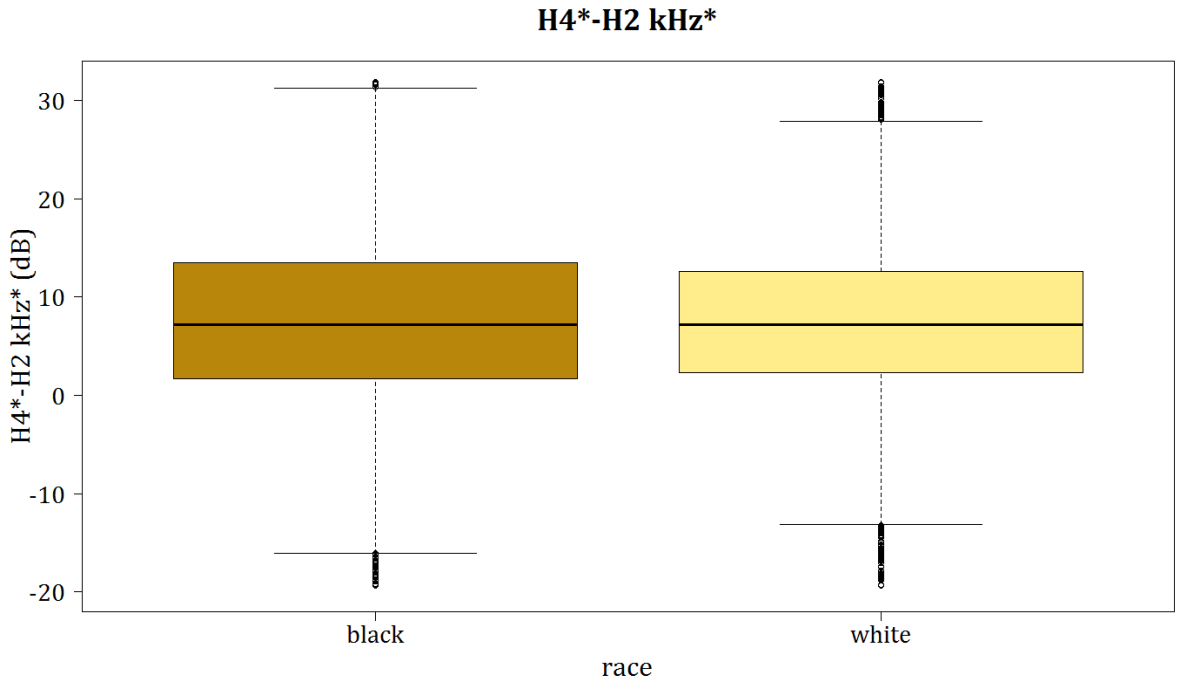

Figure 4.13: Boxplots representing the values for black and white speakers for the H4*−2K* interview data.

A similar pattern is observed for the sentence data values for this measure, as displayed in figure 4.14 below.

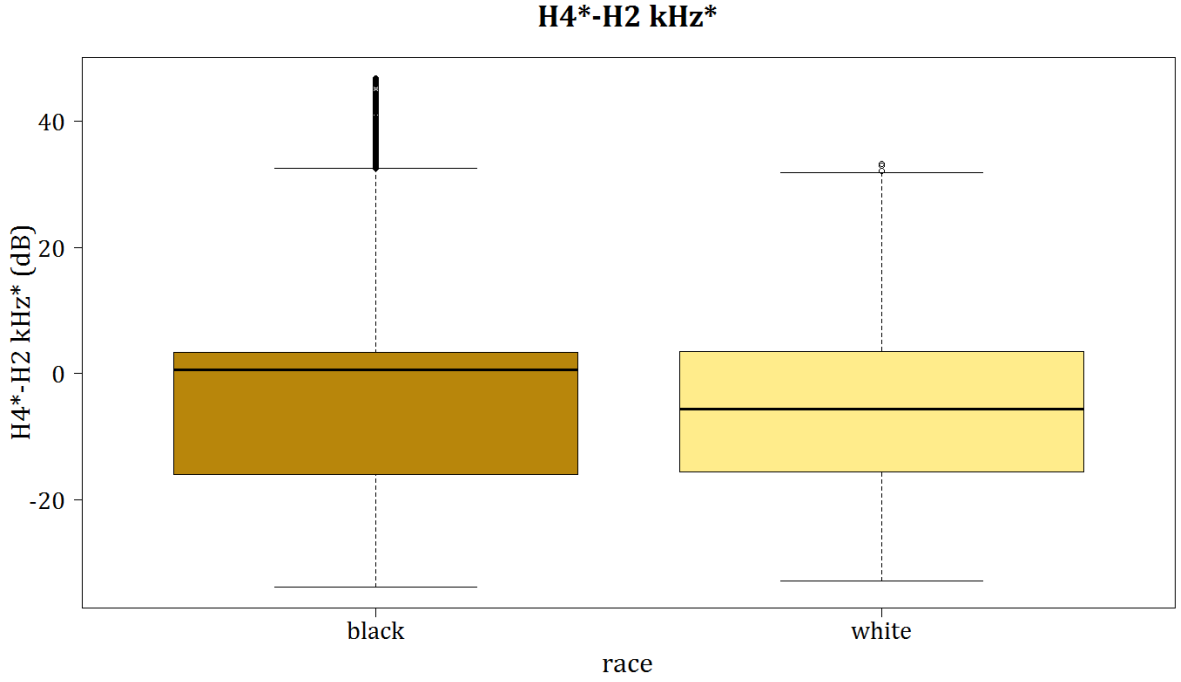

Figure 4.14: Boxplots representing the values for H4*−2K* sentence data for all data measurement points according to ethnicity.

The effect for ethnicity was found to be highly significant for the interview data for this measure according to the linear mixed effects analysis ($X^2(1)= 12.072$; $p<0.001$) with an increase of 2.196 dB ±0.57996 (standard errors) in H4*−2K* for white speakers.

The linear mixed effects analysis also revealed that the effect for logpF0 is highly significant ($X^2(1)=1808.7$; $p<0.001$) with a 1% increase in pF0 decreasing H4*−2K* by 0.093 dB ±0.021 (standard errors), as are the effects for logEnergy ($X^2(1)=9997.2$; $p<0.001$), a 1% increase in RMS Energy increasing H4*−2K* by 0.01 dB ±0.09861 (standard errors), logpF1 ($X^2(1)=536.93$; $p<0.001$), a 1% increase in pF1 increasing values by 0.073 dB ±0.30591 (standard errors), logpF2 ($X^2(1)=1776.8$; $p<0.001$), a 1% increase in duration decreases H4*−2K* by 0.013 dB ±0.33912 (standard errors), logduration ($X^2(1)=45.476$; $p<0.001$), which decreases H4*−2K* by 0.013 dB ±0.19384 (standard errors) and speaker ($X^2(3)=296.25$; $p<0.001$).

In addition, there are significant interactions for this measure between ethnicity and logpF1 ($X^2(1)= 8.7125$; $p=0.003$) and duration ($X^2(1)= 5.9182$; $p=0.015$). There are no significant interactions between ethnicity and any of the other predictors.

4.3.1.2.2.2. Sentence Data Linear Mixed Effects Analysis Results

For the sentence data, the effect for ethnicity is not significant (X^2 (1)=0.3094, $p=0.578$), although there is a highly significant effect for the interaction between ethnicity and vowel category (X^2 (2)=4943.5; $p<0.001$).

4.3.1.2.2.3. Discussion

As Garellek, Samlan, Kreiman and Gerrat (2013:3-4) observe, higher fundamental frequency is associated with a decrease in H4−2kHz[54], but lower 2kHz−5kHz and H2−H4 values are associated with an increased value for this measure, which may account for why there is no significant effect for ethnicity for the Wilcoxon rank sum tests and why there is no clear difference for the graphical comparison presented above. Black speakers display a lower fundamental frequency than white speakers, the effect of which would be to increase H4*−2K* values. However, black speakers also exhibit higher values for 2K*−5K and H2*−H4* (presented below), the effect of which would be to lower H4*−2K* values. Thus the two effects working in opposite directions may cancel out any clear difference for the boxplot comparison,

[54] The corrected measure for which is H4*−2K*.

but it would still be possible for the overall effect of ethnicity to be significant (as indicated by the results of the Wilcoxon rank sum test) and to appear as a significant predictor according to the linear mixed effects analysis.

There are also some interactions between some of the predictors and ethnicity worth further discussion. There is a significant interaction ($X^2(1)= 8.7125$; $p=0.003$) between first formant frequency and H4*−2K*. The general trend, as can be observed in figure 4.15 below, is for higher vowels, that is vowels with a lower F1, to exhibit lower values for H4*−2K*. This could be interpreted as an increase in breathiness for lower vowels.

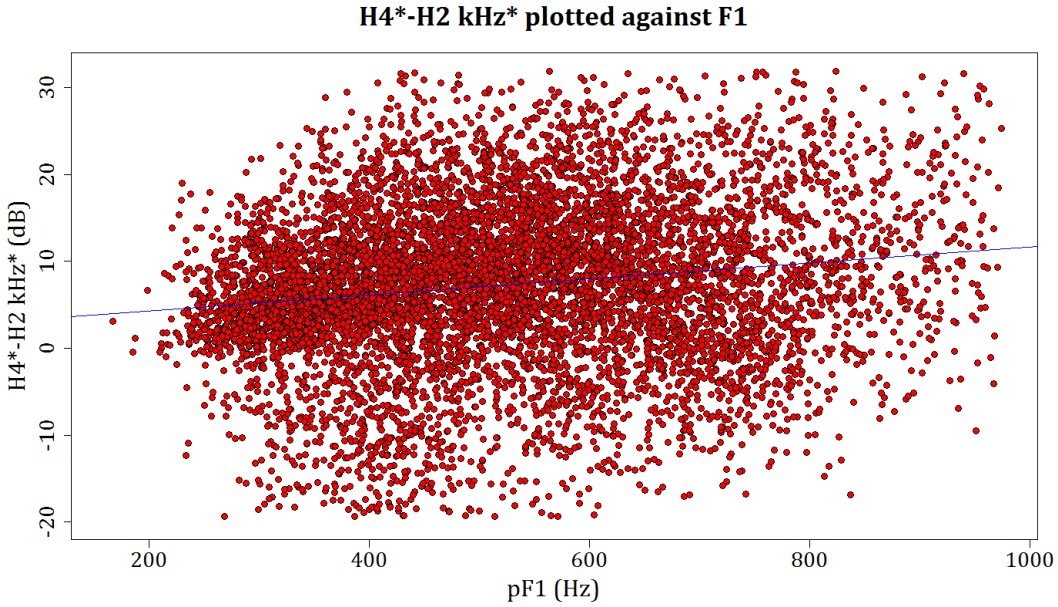

Figure 4.15: Scatterplot of the H4*−2K* data for the sample as whole plotted against pF1 in Hertz (Pearson's $r= 0.161$).

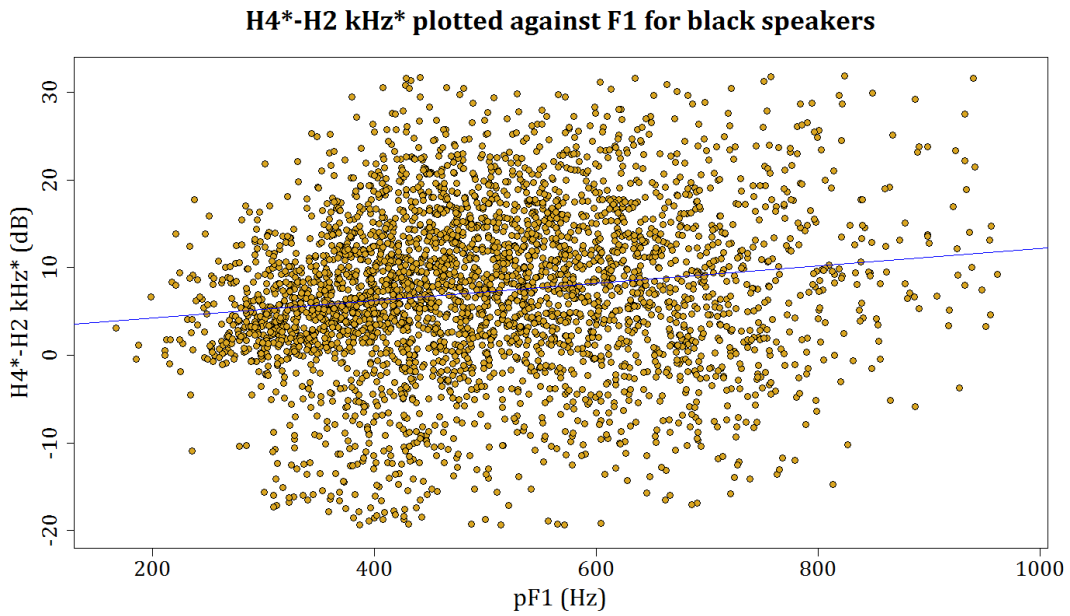

Figure 4.16: Scatterplot of the H4*−2K* data for black speakers plotted against pF1 in Hertz (Pearson's *r*= 0.152).

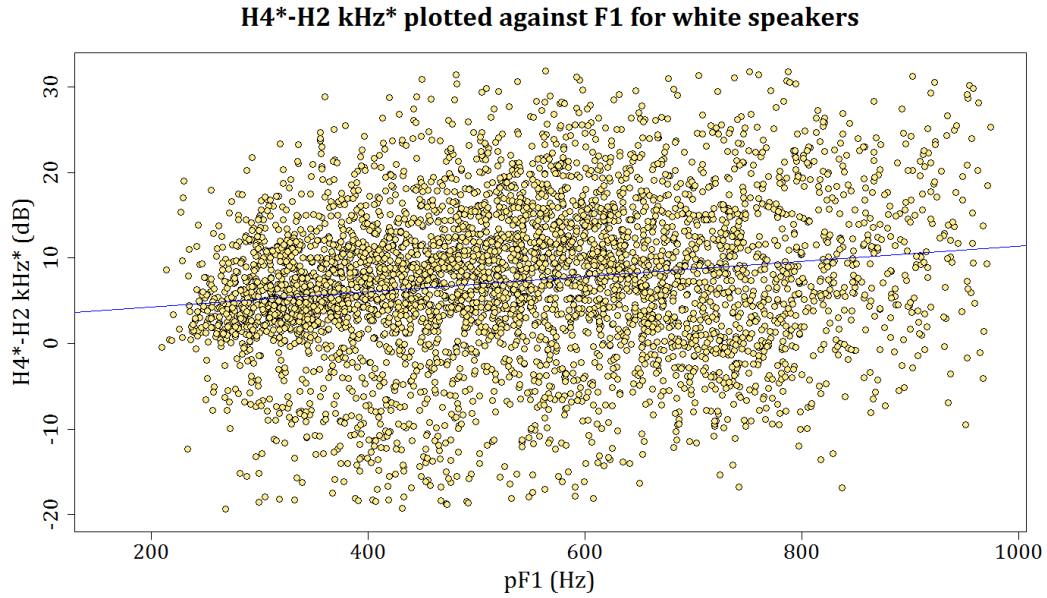

Figure 4.17: Scatterplot of the H4*−2K* data for white speakers plotted against pF1 in Hertz (Pearson's *r*= 0.171).

There is a negative correlation between this measure and duration, such that longer vowels tend to have lower values for this measure, as can be seen from figure 4.18 below ($X^2(1)= 5.9182$; $p=0.015$). This may suggest that longer vowels are less breathy. This may perhaps be due to greater use of creak for longer vowels, as opposed to breathy voice, an interpretation which would agree with the findings from the auditory analysis.

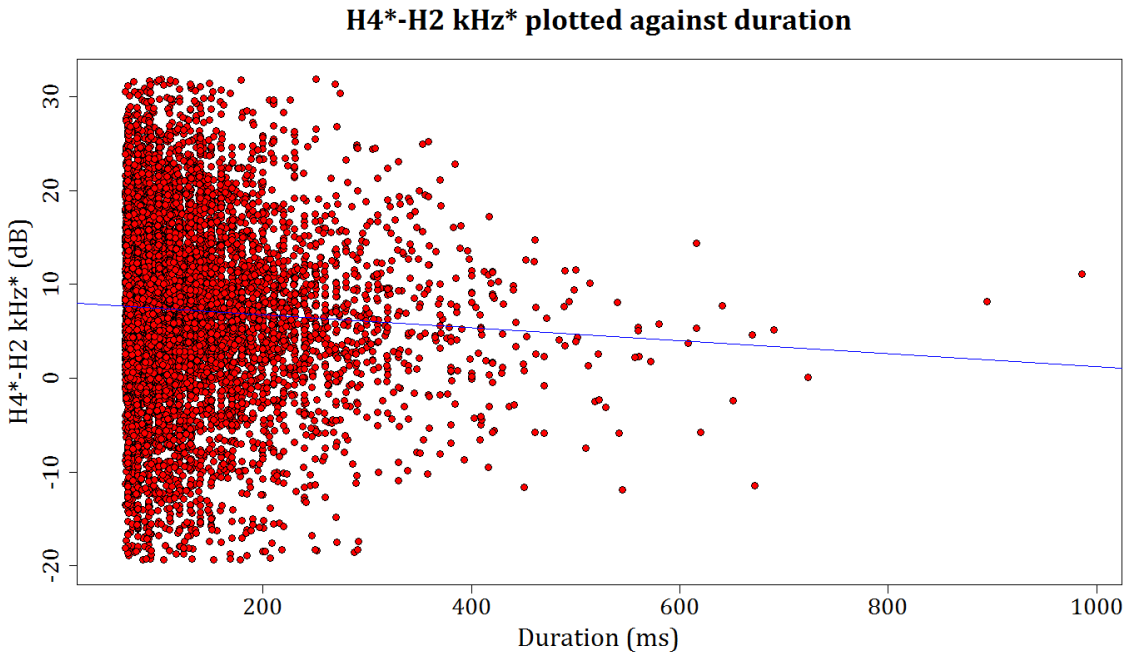

Figure 4.18: Scatterplot of H4*−2K* values for the whole sample plotted against duration in milliseconds (Pearson's $r= -0.055$).

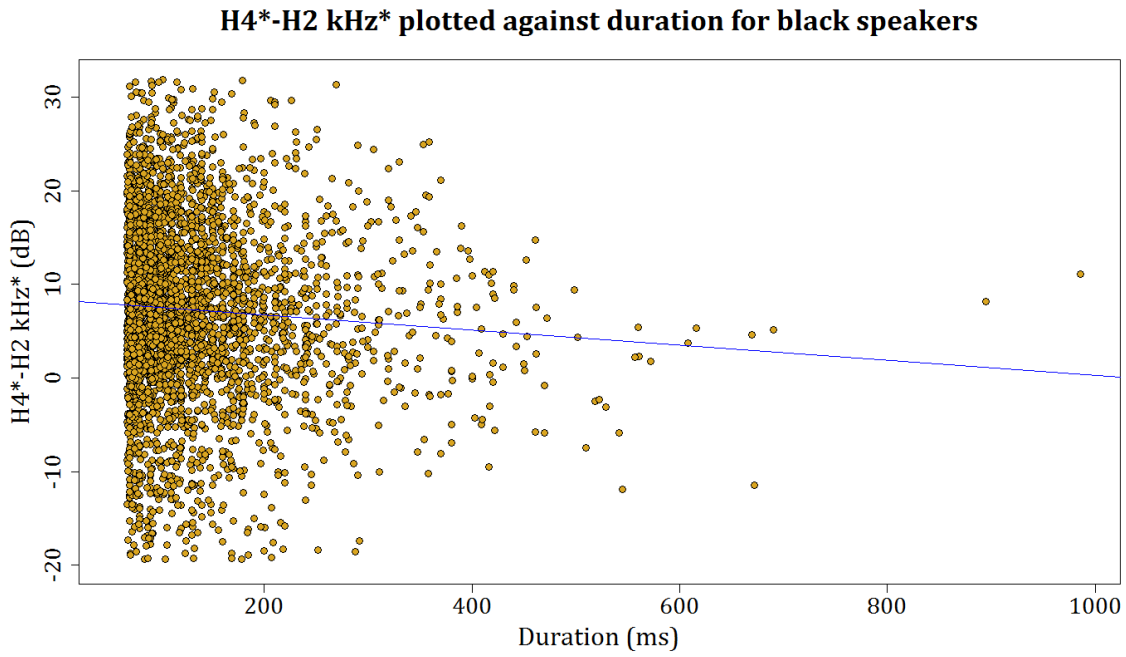

Figure 4.19: Scatterplot of H4*−2K* data for black speakers plotted against duration in milliseconds (Pearson's *r*= −0.067).

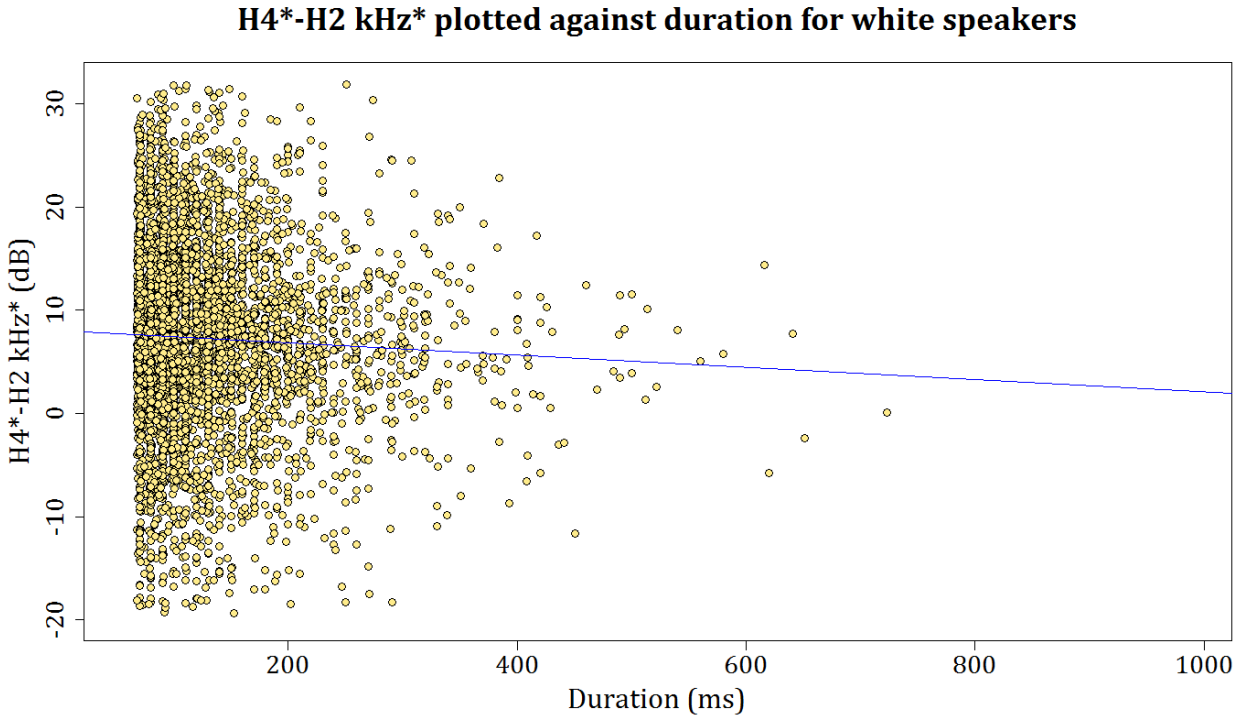

Figure 4.20: Scatterplot of H4*−2K* data for white speakers plotted against duration in milliseconds (Pearson's *r*= −0.045).

4.3.1.2.3. Summary of the Findings for H4*−2K*

The results of both the Wilcoxon rank sum tests as well as the results of the linear mixed effects regression analysis revealed a significant effect for ethnicity, with increasing values of H4*−2K* for white speakers. The effects of all other predictors were found to be significant according to the linear mixed effects analysis. In addition, there are significant interactions between ethnicity and both first formant frequency as well as duration for this measure.

As observed by Garellek, Keating, Esposito and Kreiman (2013), as well as Garellek, Samlan, Kreiman and Gerratt (2016), higher values for this measure are known to correlate with breathy voice quality, which finds support from the auditory analysis of this study. However, it may be that the effects of the other factors cancel out any clear differences in this measure between black and white speakers, such that overall no clear difference is graphically observable. Thus while there is a highly significant effect for ethnicity according to the statistical analysis, this does not appear to have an overall effect that would be detected by means of a graphical comparison. The finding that white speakers have significantly higher values for this measure is somewhat unexpected, given the aforementioned link between higher values for this measure and breathiness and the findings for the other spectral measures which indicate that black speakers have higher values (suggesting greater breathiness). While this issue is discussed in more detail in chapter 6 (particularly in section 6.2.8), it should be noted here that one possible reason for this discrepant finding is the effect that the other measures of spectral tilt included in the psychoacoustic model may have on this measure. Thus, lower values for H1−H2, H2−H4 and H2K−H5K are expected to result in higher values for H4−2K (Garellek et al. 2016). Since white speakers do indeed exhibit lower values for H1*−H2*, H2*−H4* and H2K*−H5K, it is perhaps unsurprising that they exhibit higher values for H4K*−H2K*, as observed here. Nevertheless, this finding suggests that it would be worthwhile to test the perceptual importance of this acoustic measure in particular for South African listeners in future research.

4.3.1.2.4. Summary of results for H4−2K (the uncorrected equivalent of H4*−2K*)

The results of the Wilcoxon rank sum test for the interview data for this measure revealed a significant difference between black and white speakers, with black speakers showing evidence of higher H4−2K values overall. According to the linear mixed effects analysis of the interview

data however, there is no significant effect for ethnicity with an increase in H4−2K values for white speakers. This differs from the findings for the linear mixed effects model analysis for the sentence data, which revealed a highly significant effect for ethnicity. For the interview data, significant effects for all other predictors were found using the linear mixed effects analysis. Significant interactions were found between ethnicity and the other predictors, including fundamental frequency, RMS Energy, first formant frequency, second formant frequency as well as duration. Given that there are so many significant interactions between ethnicity and the other predictors and that there are significant effects for all of these predictors, the lack of a significant effect for this measure for the interview data is perhaps unsurprising, since the effects of formants and their bandwidths are clearly considerable for the uncorrected measure. The pattern of white speakers having higher values according to the linear mixed effects regression analysis is however in agreement with the findings for the corrected measure.

4.3.1.3. H1−H2* (The first harmonic minus the second harmonic, both corrected for the influence of formants and their amplitudes)*

The measure H1−H2 (in the form of the corrected measure H1*−H2* and its uncorrected equivalent H1−H2 in this study) is one of the more well-known and reasonably well-established measures used for assessing differences in voice quality and phonation specifically, as discussed in detail in chapter two.

4.3.1.3.1. H1*−H2* Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 4.21, illustrates the boxplots representing the values for the VS-derived harmonic differential measure H1*−H2*, according to the auditorily identified phonation types for the sentence data. From examining figure 4.21 below, it is clear that one of the lower boxplots (particularly in terms of the interquartile range and the median value) is that of prototypical creak and other creak types such as vocal fry, harsh/aperiodic voice and whispery creak, while a higher interquartile range is found for breathy voice and whisper, which have higher median and mean values. Modal voice does not appear to be as clearly distinguished from breathy voice as it is from more creaky phonation types, although this may perhaps be due to the particular coding procedure used wherein a number of tokens varying in the exact degree of breathiness would be expected to have been included within the modal category.

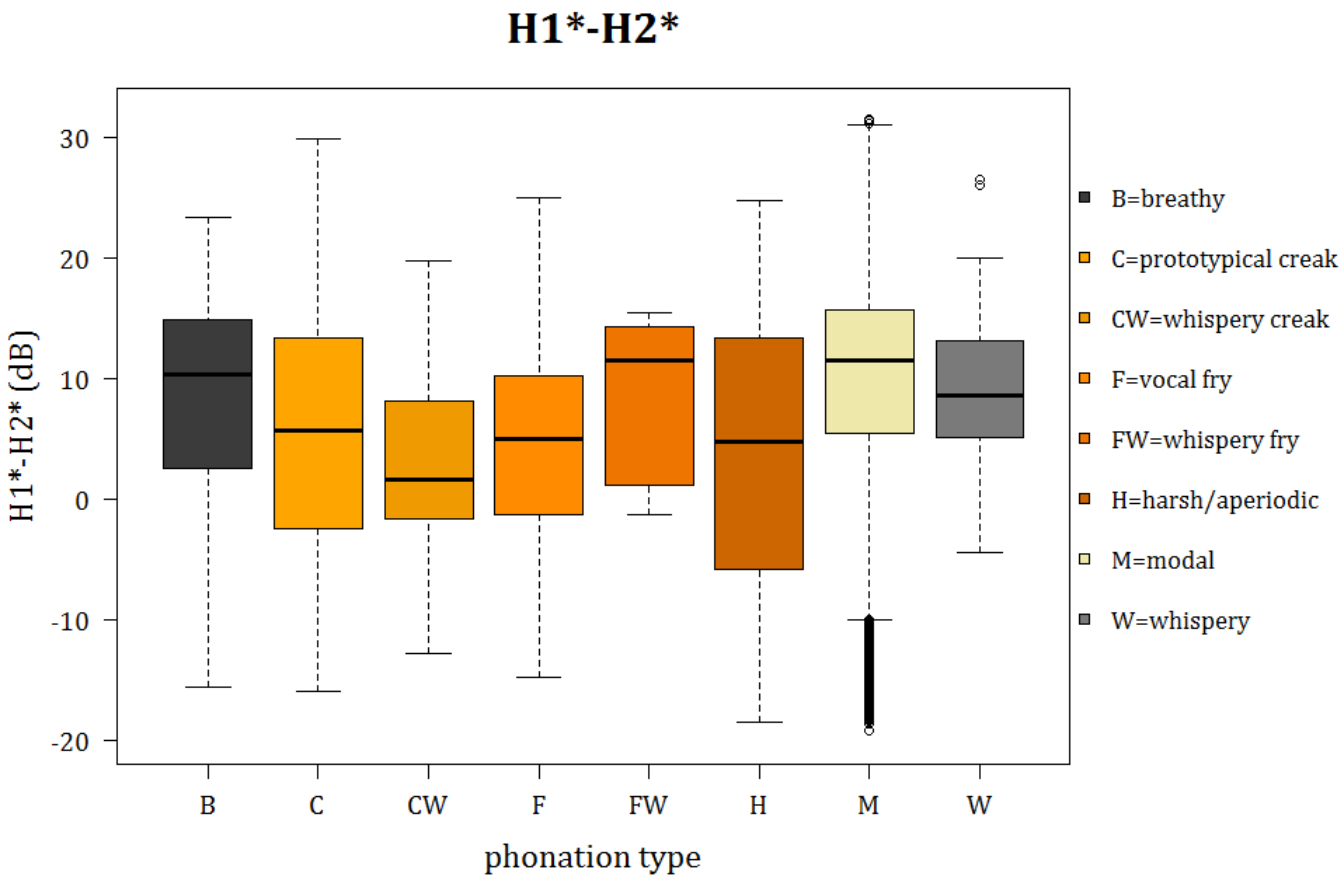

Figure 4.21: Boxplots displaying the values for the H1*–H2* sentence data for each of the auditorily identified phonation types for all data measurement points.

4.3.1.3.2. Statistical Analysis for H1*–H2*

4.3.1.3.2.1. Interview Data

The Wilcoxon rank test results do not show a significant difference between the two ethnic groups for this measure (W=203, p=0.169), for the alternative hypothesis that the values for black speakers are higher than those for white speakers. However, some difference in this direction is evident in the boxplot comparison presented in figure 4.22 below.

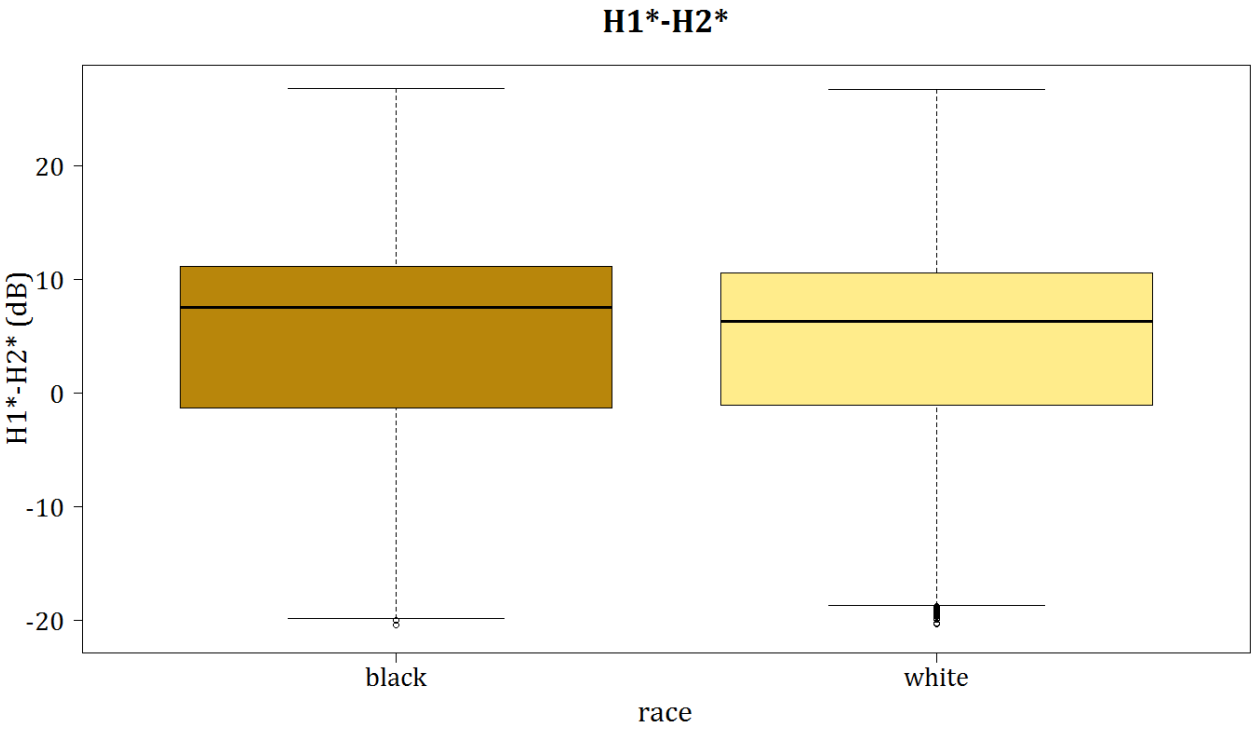

Figure 4.22: Boxplots representing the values of black and white speakers for the H1*–H2* interview data.

This pattern can also be observed for the sentence data for this measure, displayed in figure 4.23 below, although it is not as clear as for the interview data.

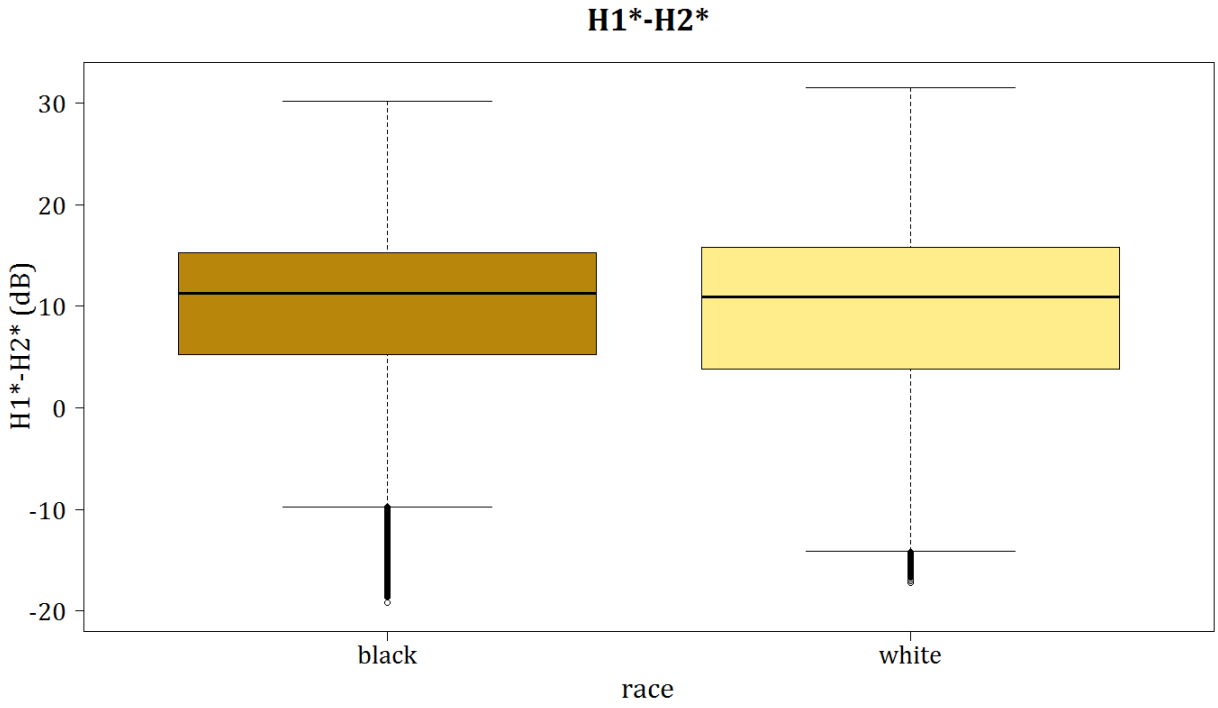

Figure 4.23: Boxplots representing the values for H1*−H2* sentence data for all data measurement points according to ethnicity.

The result for ethnicity for the linear mixed effects models regression analysis of the interview data reveals a significant effect ($X^2(1)= 6.2559$;$p=0.012$) with a decrease of 1.737 dB ±0.63568 (standard errors) in H1*−H2* values for white speakers.

There are also significant effects for most of the other predictors. Thus there is a significant effect for logpF0 ($X^2(1)=7185$;$p<0.001$) such that a 1% increase in pF0 increases H1*−H2* by 0.199 dB ±0.18415 (standard errors), logEnergy ($X^2(1)=109.23$;$p<0.001$) such that a 1% increase in RMS Energy decreases values by 0.009 dB ±0.08749 (standard errors), a nearly significant effect for logpF1 ($X^2(1)=3.4507$;$p=0.063$), a 1% increase in pF1 increasing values by 0.005 dB ±0.25464 (standard errors), logpF2 ($X^2(1)=40.84$;$p<0.001$) with a 1% increase in pF2 decreasing values by 0.018 dB ±0.27856 (standard errors) and speaker ($X^2(3)=523.38$;$p<0.001$). Logduration is the only predictor for which the effect is neither significant, nor approaching significance ($X^2(1)=0.2976$;$p=0.585$).

There are also significant interactions between ethnicity and several of the other predictors for this measure for the interview data. There is a significant interaction between ethnicity and logpF0 ($X^2(1)= 101$;$p<0.001$), logpF1 ($X^2(1)= 30.066$;$p<0.001$) and logduration ($X^2(1)= 5.3037$;$p=0.021$). The interaction between logpF2 and ethnicity approaches significance ($X^2(1)= 2.8494$;$p=0.091$), as does the interaction between logEnergy and ethnicity ($X^2(1)= 3.3658$;$p=0.067$).

4.3.1.3.2.2. Sentence Data Linear Mixed Effects Analysis Results

For the sentence data I found no significant effect for ethnicity (X^2 (1)=0. 4403, $p=0.507$), but a highly significant effect for the interaction between vowel and ethnicity (X^2 (2)=264.13, $p<0.001$).

4.3.1.3.2.3. Discussion

For the interview data, that is, the data which most directly pertains to voice quality, the pattern observed and the fact that there is a significant difference for this measure would suggest that there is overall greater glottal stricture for white speakers in comparison to black speakers (assuming no differences in glottal chink) and a smaller open quotient (OQ) for white speakers, which would usually suggest that black speakers may exhibit greater aspiration noise and potentially use more breathy voice, while white speakers may tend towards the use of more modal or creaky phonation types. This finding parallels that of Szakay and Torgersen (2015), who concluded that inner-city non-Anglo London English speakers have breathier phonation than Anglo speakers who live on the periphery of London.

However, there are also a number of significant interactions for this measure between ethnicity and some of the other predictors, and I will therefore consider the possible contributions of these interactions in turn.

In figures 4.25, 4.26 and 4.27 below, the H1*−H2* values are plotted against the fundamental frequency measure pF0. There is a clear positive correlation between this measure and fundamental frequency for the sample as a whole and for both ethnic groups. The positive correlation between this measure and fundamental frequency is expected, as reported by

Garellek, Samlan, Kreiman and Gerratt (2013:3). The correlation is however more strongly positive for black speakers. This could suggest that the increase in breathiness or at least a decrease in creak associated with an increase in pitch is greater for black speakers than for white speakers. However, the patterning of the H1*−H2* data when plotted against F0 is more clearly bimodal for black speakers, suggesting that for black speakers, low H1*−H2* values more clearly correlate with low F0.

In spite of fundamental frequency being lower for black speakers overall, as can be seen from figure 4.24 below, which one may expect to condition lower values for H1*−H2* overall, the values are nevertheless higher for black speakers than for white speakers. This may suggest the importance of an ethnic difference. That is, in comparison to Szakay and Torgersen's (2015) findings both groups of speakers in my study are similar to the female Anglo speakers in that study (in terms of higher pitch being associated with higher H1−H2 values), however the effect is stronger for the black speakers in my study.

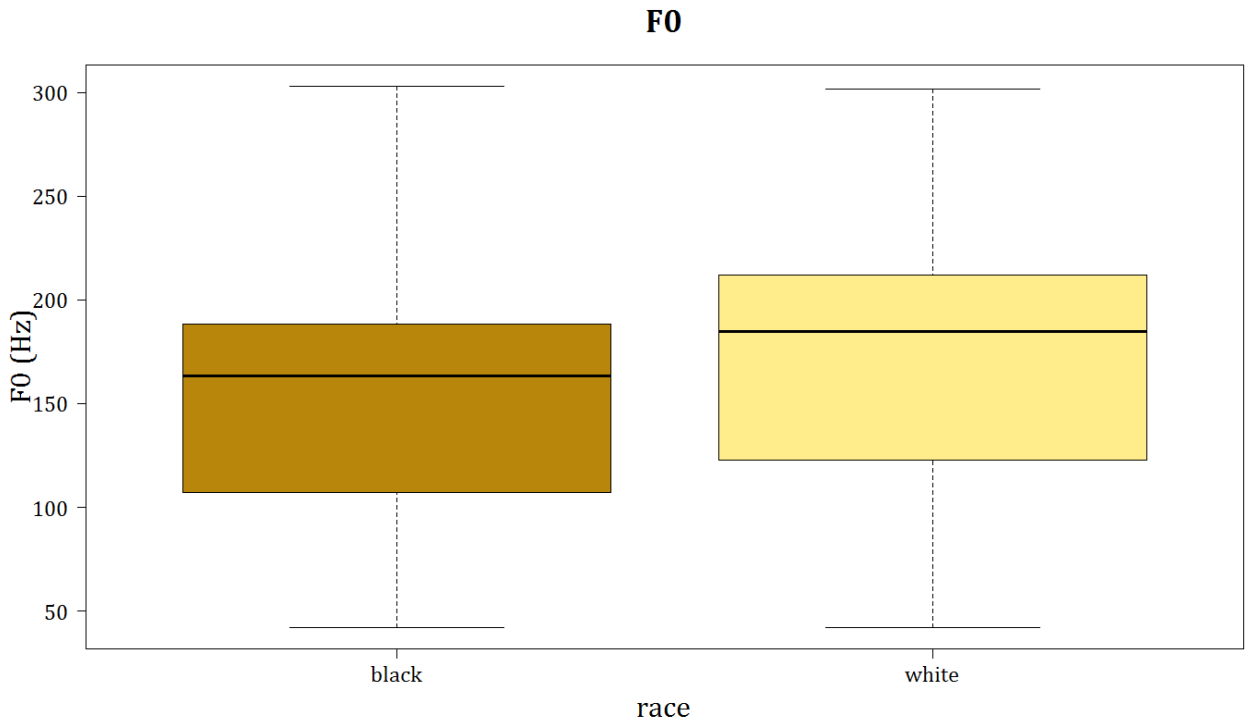

Figure 4.24: Boxplots representing the values of black and white speakers for the pF0 interview data.

That black speakers show a stronger positive correlation between fundamental frequency and this measure may in fact be attributed to the finding by Iseli, Shue and Alwan (2007) that for f_0 values below 175 Hz there is a strong positive correlation of (r=0.77), while a negative correlation obtains for fundamental frequency values above 175 Hz (r=−0.47). Given that black speakers display lower values for fundamental frequency overall, this stronger correlation than for white speakers is unsurprising. However, in spite of the lower values for fundamental frequency, black speakers overall still display higher values for H1*−H2*, which may thus provide further evidence of an ethnic difference in voice quality.

Iseli, Shue and Alwan (2006:I-390) found a 0.56 Pearson product correlation for their low pitched speakers and hypothesize that the simultaneous increase of both f_0 and H1*−H2* may be related to increased cricothyroid tension (Iseli, Shue and Alwan 2006:I-390). The findings for this measure may therefore potentially signify greater importance of cricothyroid tension during phonation for black speakers than for white speakers.

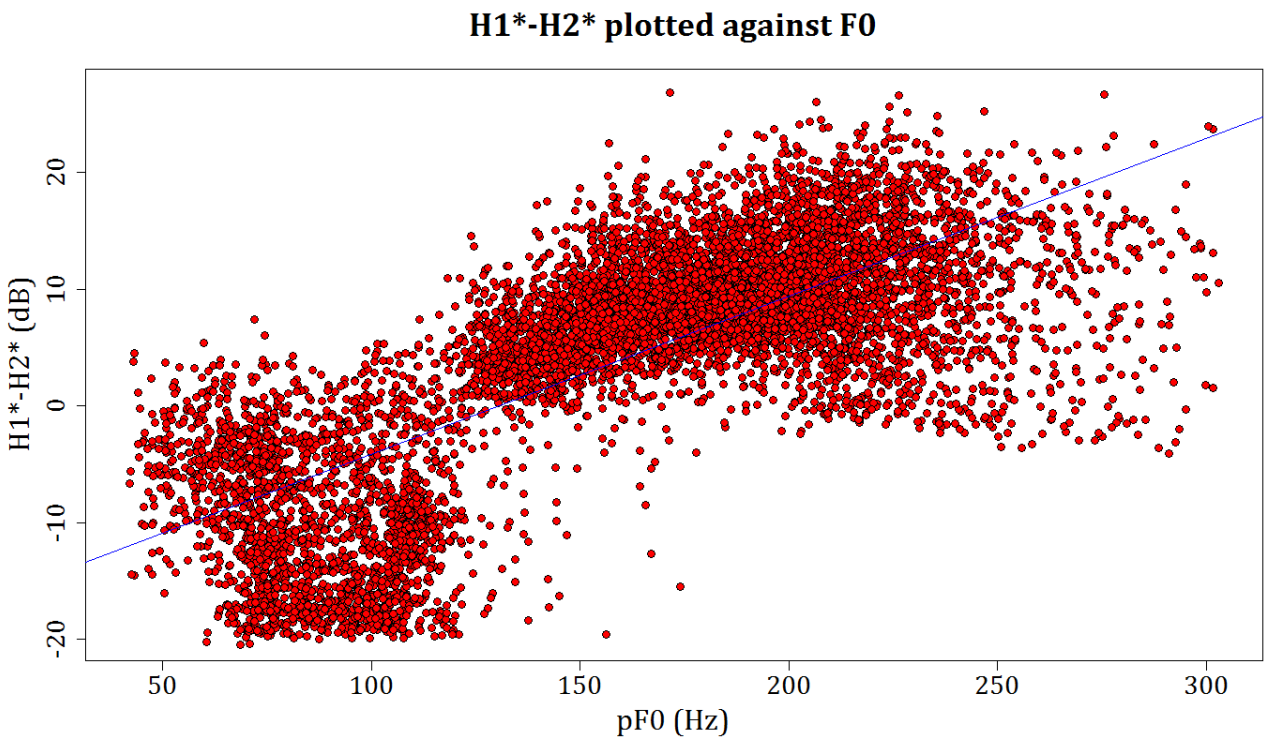

Figure 4.25: Scatterplot of H1*–H2* data for the whole sample plotted against pF0 in Hertz (Pearson's *r*= 0.750).

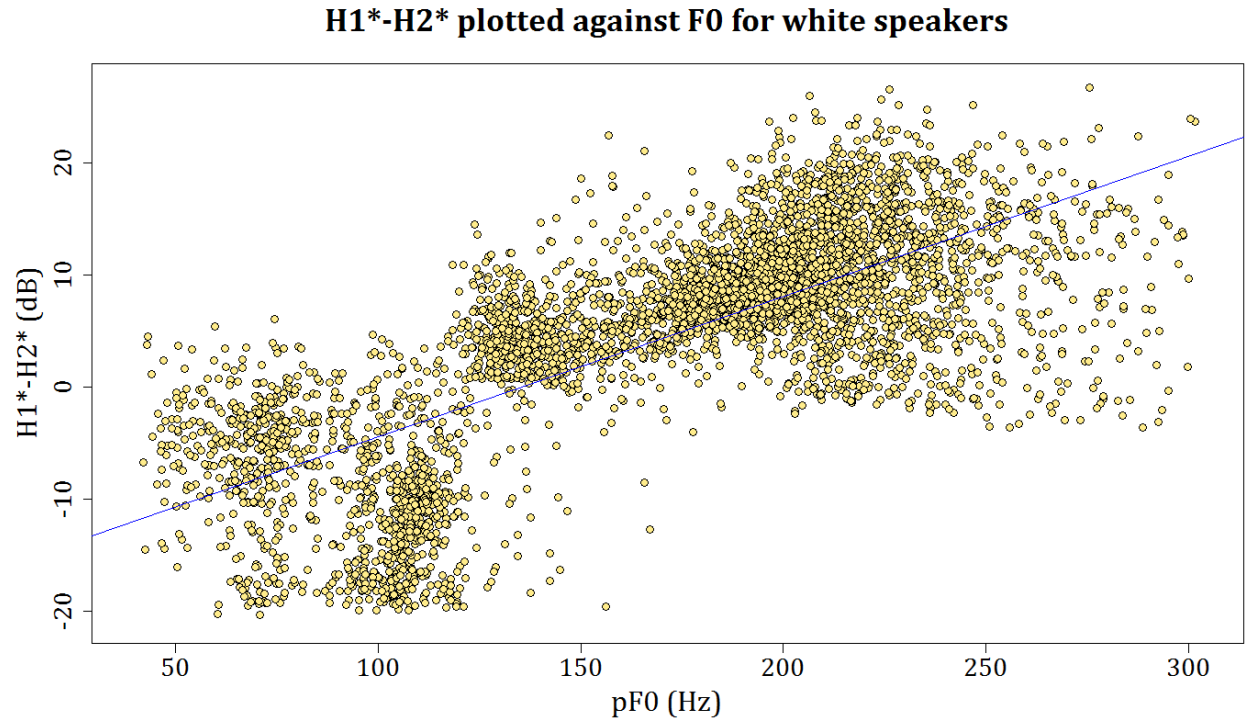

Figure 4.26: Scatterplot of H1*–H2* data for the white speakers plotted against pF0 in Hertz (Pearson's *r*= 0.742).

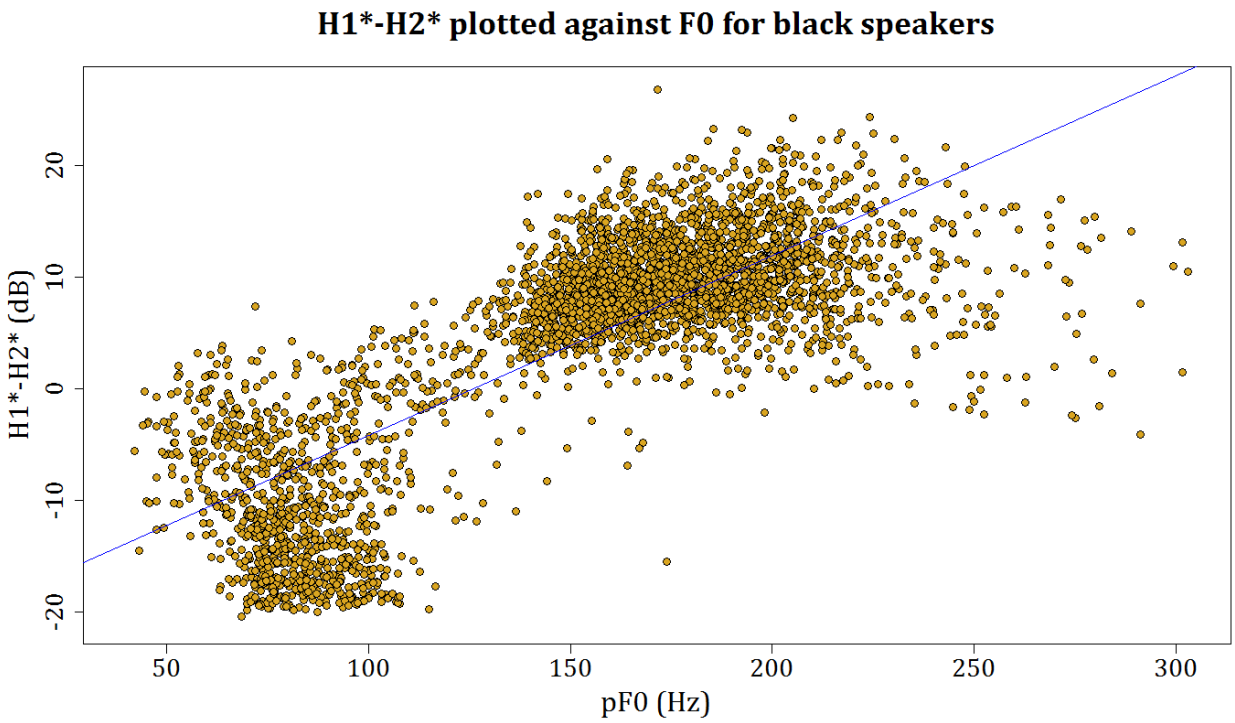

Figure 4.27: Scatterplot of H1*−H2* data for the black speakers plotted against pF0 in Hertz (Pearson's *r*= 0.790).

A comparison of the scatterplots provided in appendix E reveal that, for the interview data of the entire sample, there is no obvious correlation between H1*−H2* and first formant frequency. Black speakers exhibit a weak negative correlation and white speakers a weak positive correlation.

The findings for the white speakers in my study as they relate to the interaction between this measure and first formant frequency are therefore similar to the finding of Szakay and Torgersen (2015) for their sample overall.

The findings for white speakers in my study are also similar to those of Iseli, Shue and Alwan (2006:I-391) for their high pitched talkers, where the Pearson product correlation was found to be 0.61 although the correlation in my study is far weaker (Pearson's *r*=0.028) and H1*−H2* values were found to increase along with first formant frequency for frequencies under 700Hz.

Iseli, Shue and Alwan (2007) reported that none of their groups of speakers showed a correlation between H1*−H2* and F2. In my data, as evidenced by a comparison of the scatterplots (and Pearson's *r* values) provided in appendix E, the correlation is likewise so weak as to be negligible for the sample as a whole as well as for the ethnic subsamples.

For duration, as illustrated by the scatterplots presented in appendix E, there is a weak negative correlation between H1*−H2* and duration. This is not surprising, given that Gerratt, Kreiman and Garellek (2016:5) found that the mean is somewhat lower for continuous vowels (generally shorter) in comparison to sustained vowels (generally longer) for this measure. The effect is clearer for black speakers than for white speakers.

Szakay and Torgersen (2015) found significantly higher values for H1−H2 for longer vowels but only for male speakers. For my data from all female speakers, there is a correlation with duration, however it is weakly negative rather than positive and stronger for black speakers than for white speakers.

Regarding the relationship between intensity and H1−H2, Szakay and Torgersen (2015) found a decrease for this measure associated with increased intensity. For the RMS energy measure used in my study, I found a weak relationship in the opposite direction (as can be seen from the scatterplots provided in appendix E), however this is difficult to interpret because the distributions for this measure when plotted against RMS energy appear to be bimodal. Chen, Park, Kreiman and Alwan (2014) also found that for some of their speakers, there was a positive correlation between intensity and H1*−H2*.

4.3.1.3.3. Summary of Findings for H1*−H2*

The results of the Wilcoxon rank sum tests revealed no significant difference between black and white speakers for this measure, although higher values overall for black speakers are evident from the boxplot comparison. There is however a significant effect for ethnicity according to the results of the linear mixed effects analysis for this measure, with the effect being a decrease in H1*−H2* values for white speakers. All other predictors apart from duration were found to have significant effects according to the linear mixed effects analysis. Significant interactions were found between ethnicity and several of the other predictors, including fundamental frequency, first formant frequency as well as duration. Effects approaching significance were found for the

interactions between ethnicity and second formant frequency as well as between ethnicity and RMS Energy.

Keating and Esposito (2006) regard this measure as a useful indicator of overall stricture of the glottis along the glottal stricture continuum due to the relationship between OQ (see footnote on page 55) and such stricture. As pointed out by Chen, Park, Kreiman and Alwan (2014) however, more recent research indicates that there are other factors which influence H1−H2 values including fundamental frequency, glottal pulse skewness as well as the glottal gap. Following Keating et al. (2015), this measure reflects degree of glottal constriction with lower values for this measure suggesting greater glottal constriction, although assuming that glottal chink does not predominate, this measure also provides a relatively accurate reflection of open quotient. Thus the overall results of my study would indicate that glottal constriction is greater for white speakers than for black speakers and possibly that open quotient is greater for black speakers. These findings may therefore also suggest that black speakers use a voice quality which involves more aspiration than white speakers.

4.3.1.4. H1–H2 (the uncorrected equivalent measure of H1–H2*)*

While I have not included details regarding the uncorrected measures for most of the harmonic amplitude measures in the main text of this thesis (I have included basic descriptive statistics for these measures in appendix C), for comparative purposes, given that numerous earlier studies have used the uncorrected measure H1−H2, I have included the findings for this measure in the following section.

4.3.1.4.1. H1–H2 Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 4.28 illustrates the boxplots representing the values for the H1−H2 sentence data, according to the auditorily identified phonation types for all data measurement points. The highest overall interquartile range is that of the breathy voice category as would be expected given that the previously described correlations for this auditorily identified phonation type also exhibits one of the most negatively skewed distributions of all phonation types. This is expected, given that phonation types such as breathy voice and whisper are expected to exhibit higher values for H1−H2 than modal voice for example. The boxplots generally correspond to the expected values for the phonation types, as described in the literature.

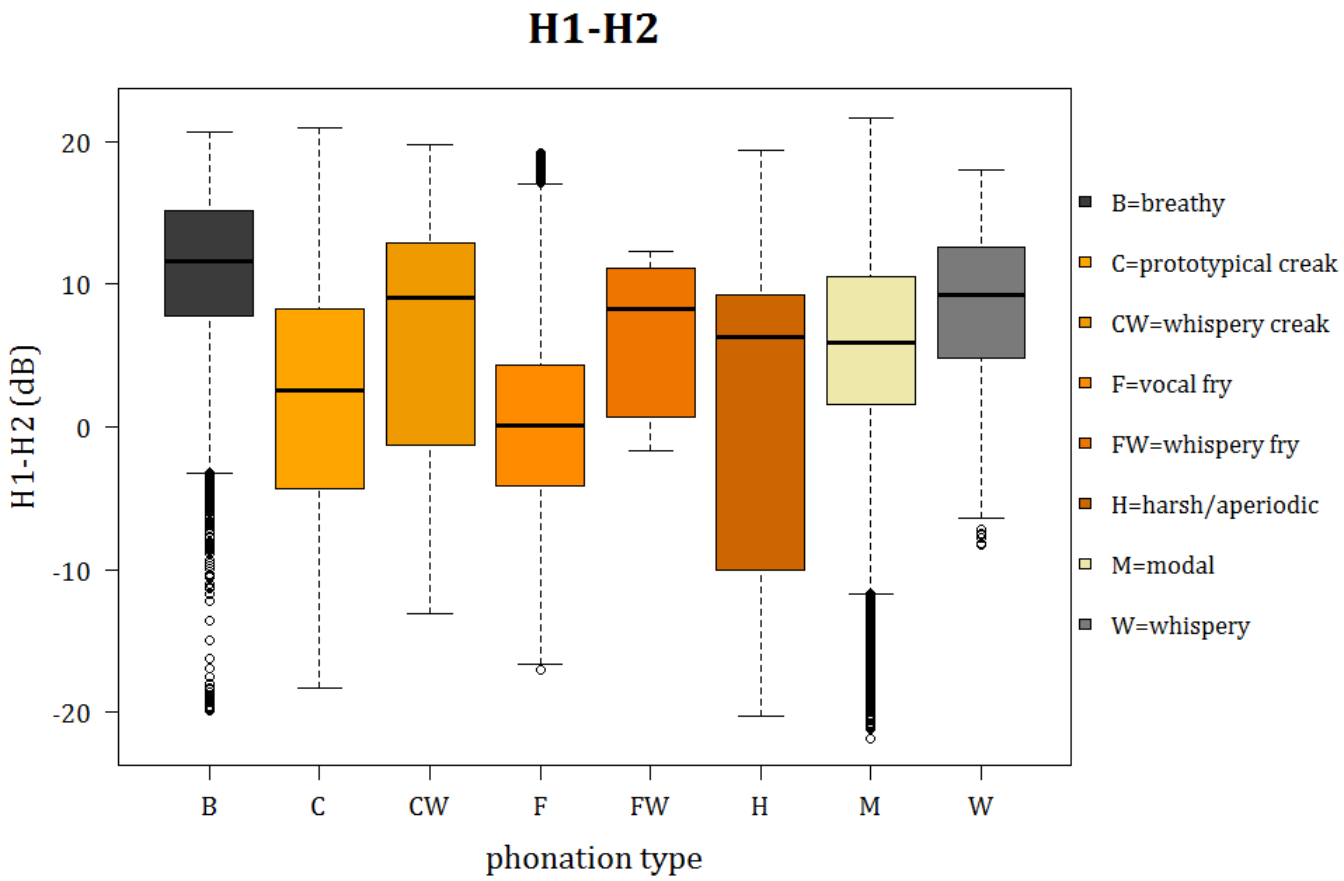

Figure 4.28: Boxplots displaying the values for the H1−H2 sentence data for each of the auditorily identified phonation types for all data measurement points.

The lowest values are generally found for the non-compound creak types, such as prototypical creak, vocal fry and harsh/aperiodic. The boxplot for modal voice appears to fall between the extremes of breathy voice on the one hand and vocal fry on the other.

4.3.1.4.2. Statistical Analysis for H1−H2

4.3.1.4.2.1. Interview Data

The results of the Wilcoxon rank sum test for the H1−H2 measure reveal no significant effect for ethnicity (W=183, p=0.261). There is no particularly clear difference between black and white speakers for this measure, as can be seen from figure 4.29 below, however the median value for white speakers is lower than for black speakers. As illustrated in appendix C however, the mean value is higher for white speakers (1.053 dB) and lower for black speakers (0.133 dB).

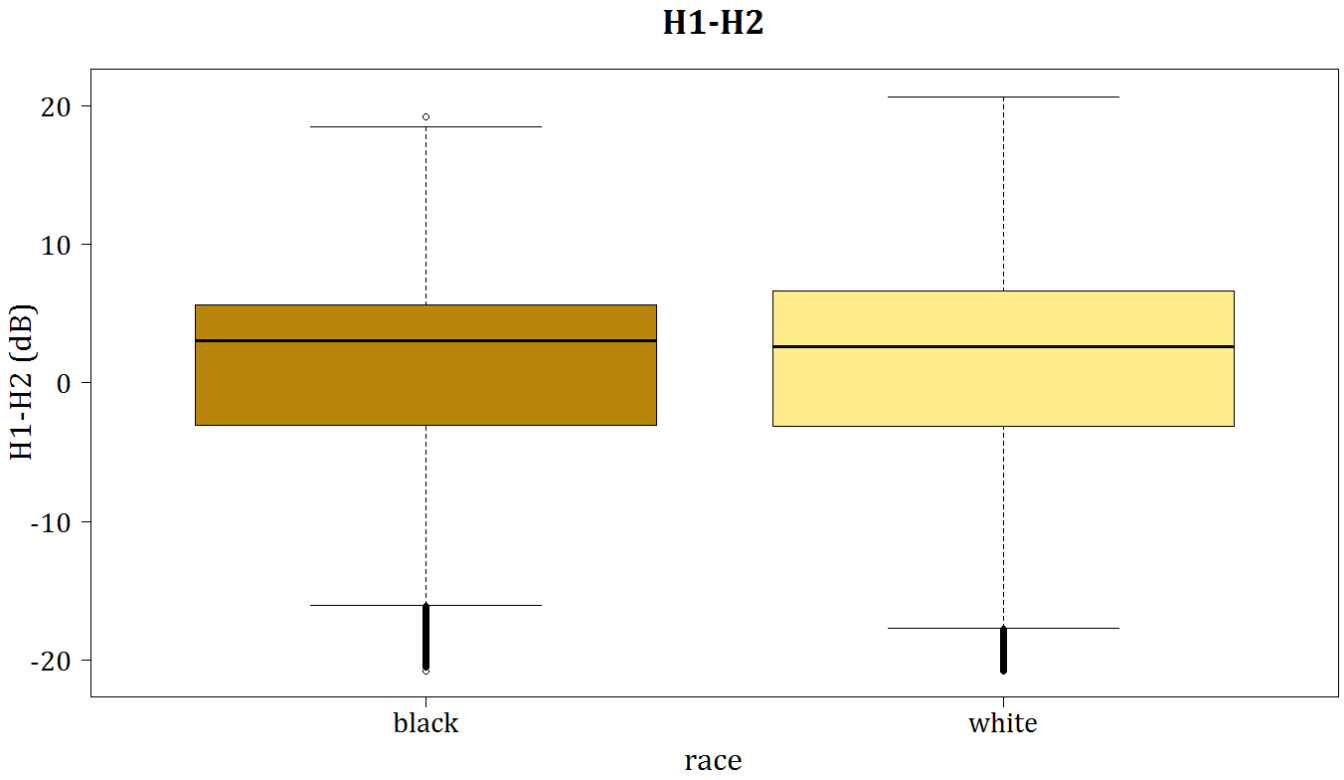

Figure 4.29: Boxplots representing the values for black and white speakers for the H1−H2 interview data.

The pattern is similar for the sentence data, displayed in figure 4.30 below.

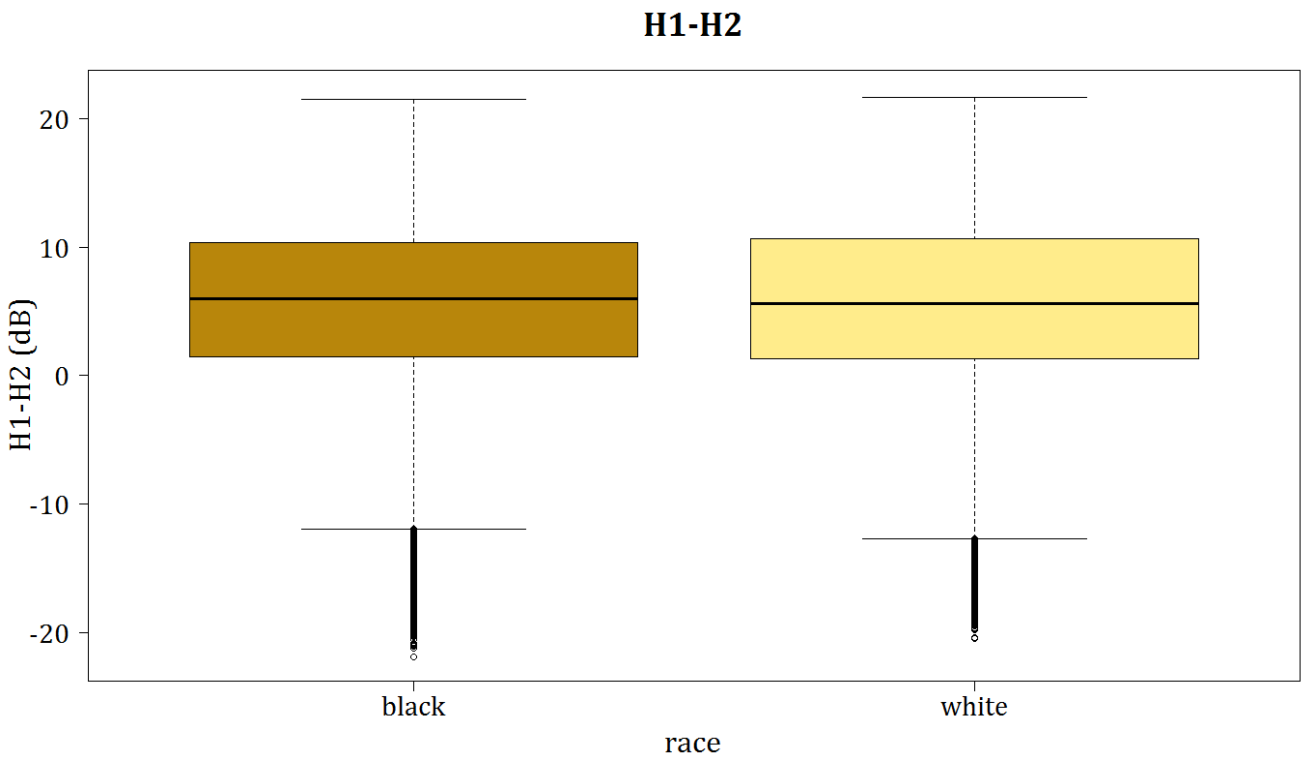

Figure 4.30: Boxplots representing the values for H1−H2 sentence data for all data measurement points according to ethnicity.

I found no significant effect for ethnicity for the H1−H2 interview data according to the linear mixed effects analysis ($X^2(1)= 1.3422$;$p=0.247$) with a decrease in H1−H2 values of 0.979 dB ±0.83312 (standard errors) for white speakers.

There is however, a highly significant effect for logpF0 ($X^2(1)=7438.2$;$p<0.001$) such that a 1% increase in pF0 increases H1−H2 by 0.176 dB ±0.15846 (standard errors), logEnergy ($X^2(1)=936.38$;$p<0.001$), a 1% increase in RMS Energy decreasing H1−H2 by 0.024 dB ±0.07581 (standard errors), logpF1 ($X^2(1)=28.59$;$p<0.001$), a 1% increase in pF1 increasing values by 0.012 dB ±0.21648 (standard errors) and speaker ($X^2(3)=926.92$;$p<0.001$). There is also a significant effect for logpF2 ($X^2(1)=4.8008$;$p=0.028$), a 1% increase in pF2 decreasing values by 0.005 dB ±0.23635 (standard errors) while the effect for duration is not significant for this measure ($X^2(1)=1.9264$;$p=0.165$).

There is a highly significant interaction between ethnicity and logpF0 ($X^2(1)=$ 21.554;$p<0.001$) for the interview data, while the interaction between ethnicity and logpF1

approaches significance ($X^2(1)= 3.3067$;p=0.069). There are no other significant interactions between the other predictors and ethnicity.

4.3.1.4.2.2. Sentence Data Linear Mixed Effects Analysis Results

As for the interview data, there is no significant effect for ethnicity (X^2 (1)=0.0808;p=0.776) for the sentence data. However, there is a highly significant interaction between ethnicity and vowel (X^2 (2)=675.25, p<0.001).

4.3.1.4.3. Summary of Findings for H1−H2

There is no significant difference between black and white speakers according to the Wilcoxon rank sum test results. According to the linear mixed effects analysis however, there is no significant effect for ethnicity for the uncorrected measure, contrary to the findings for the corrected measure. All other predictors were found to have significant effects according to the linear mixed effects analysis, apart from duration. There is also a significant interaction between ethnicity and fundamental frequency and an interaction approaching significance between ethnicity and first formant frequency.

4.3.1.5. H2−H4* (the second harmonic minus the fourth harmonic, both corrected for the influence of formants and their amplitudes)*

4.3.1.5.1. H2*−H4* Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 4.31, displays the boxplots for the H2*−H4* sentence data grouped according to auditorily identified phonation types for all data measurement points. In general, it appears that for this measure, the boxplots for the different phonation types are not very clearly distinguished from one another. The widest range (both in terms of the overall range as well as the interquartile range) is found for vocal fry, suggesting that there is a high degree of variability for this category for the measure H2*−H4*. Similarly wider ranges are found for other non-compound creak types such as prototypical creak and harsh/aperiodic voice. Higher medians are generally found for creak types.

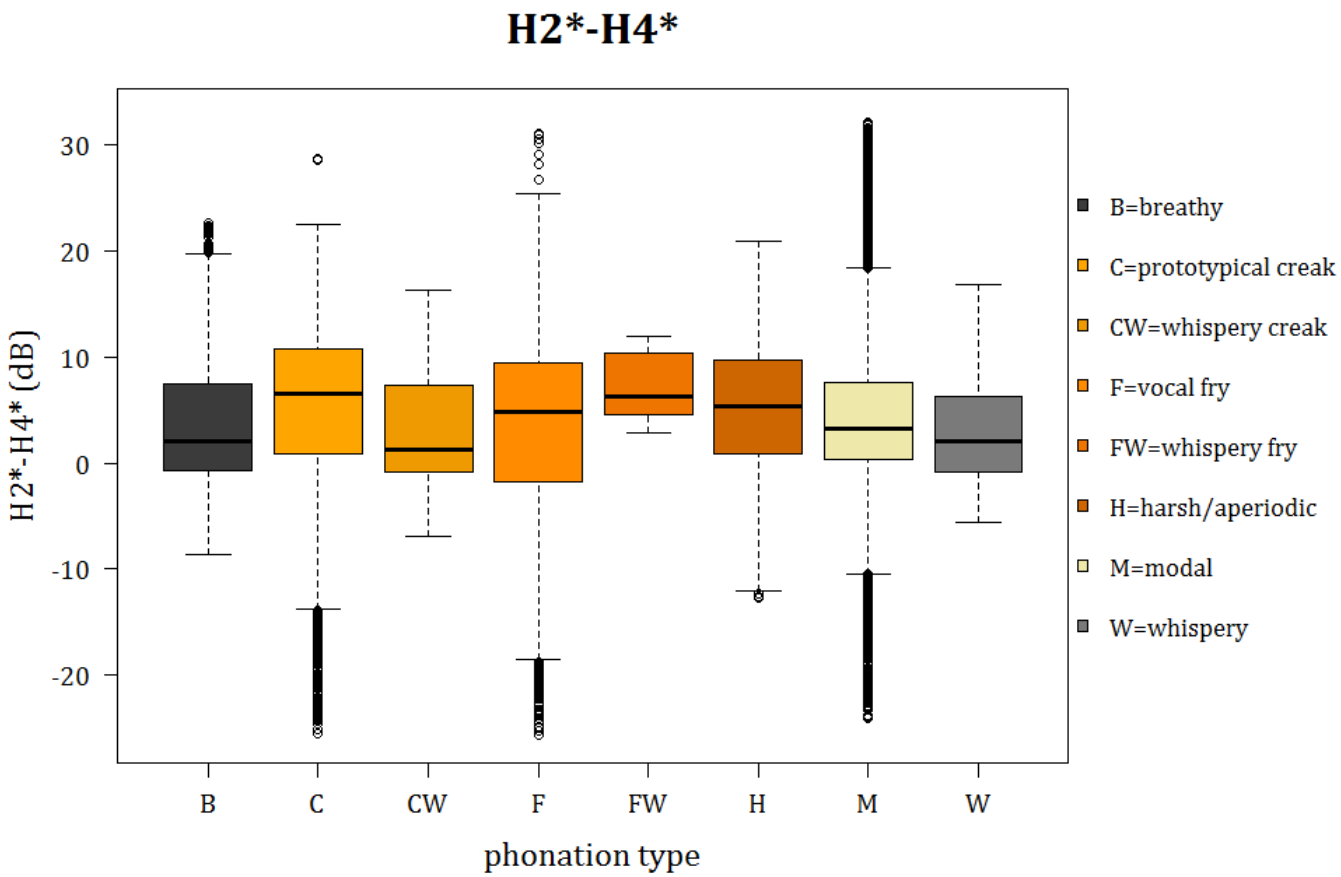

Figure 4.31: Boxplots displaying the values for the H2*−H4* sentence data for each of the auditorily identified phonation types for all data measurement points.

4.3.1.5.2. Statistical Analysis for H2*−H4*

The results of the Wilcoxon rank sum test for this measure reveal that the two groups are not significantly different (W=200, *p*=0.119). There does however appear to be an overall difference as is evident from the boxplot comparison presented in figure 4.32 below, where black speakers appear to have higher values overall.

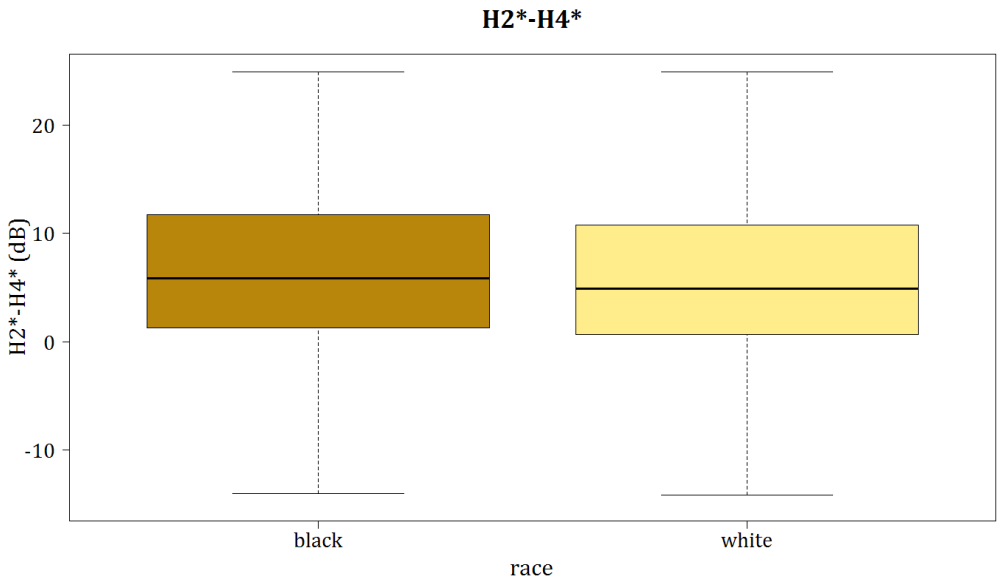

Figure 4.32: Boxplot comparison for the values of white and black speakers for the H2*−H4* interview data.

There is little distinguishing the two groups for the sentence data, as can be seen from figure 4.33 below.

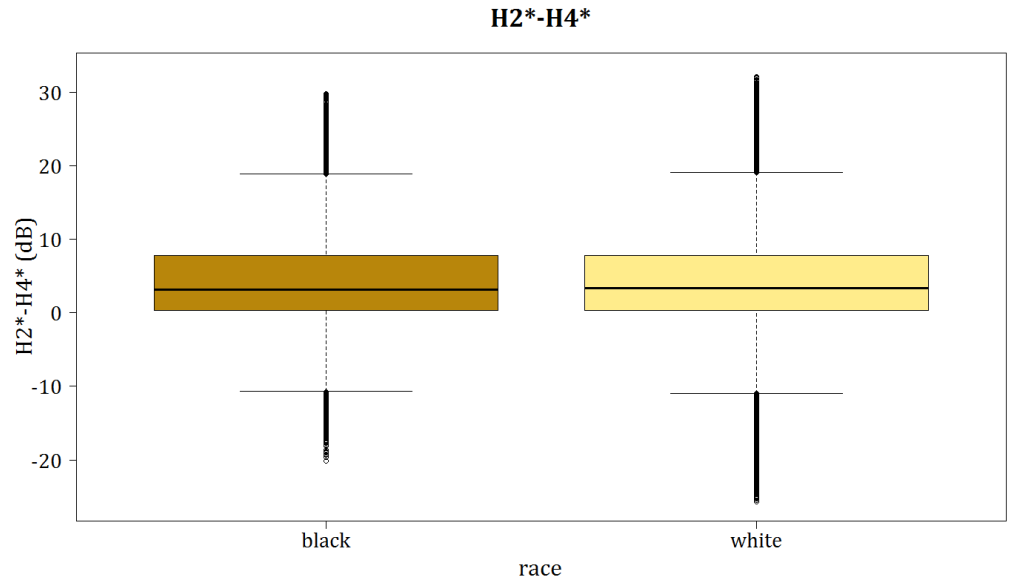

Figure 4.33: Boxplots representing the values for H2*−H4* sentence data for all data measurement points according to ethnicity.

4.3.1.5.2.1. Interview Data

The results of the linear mixed model regression analysis reveal that there is no significant effect for ethnicity for this variable for the interview data ($X^2(1)= 2.41$;$p=0.121$) with a decrease of 0.86 dB ±0.54809 (standard errors) in the values for H2*−H4* for white speakers.

However, there are significant effects and effects approaching significance for several of the other predictors as well as significant effects for the interaction between ethnicity and some of these predictors for the interview data. There is a significant effect for logpF0 ($X^2(1)=405.95$;$p<0.001$), a 1% increase in pF0 decreasing H2*−H4* values by 0.041 dB ±0.19921 (standard errors), a significant effect for logpF1 ($X^2(1)=1080.8$;$p<0.001$), a 1% increase in pF1 increasing values by 0.093 dB ±0.27206 (standard errors), for logduration ($X^2(1)=12.967$;$p<0.001$), a 1% increase in duration decreasing values by 0.007 dB ±0.18205 (standard errors) and a significant effect for speaker ($X^2(3)=199.84$;$p<0.001$). The effect for logpF2 approaches significance ($X^2(1)=3.0221$;$p=0.082$), a 1% increase in pF2 decreasing values by 0.005 dB ±0.29722 (standard errors), while logEnergy is not significant ($X^2(1)=1.4753$;$p=0.225$).

There are also highly significant interactions between ethnicity and logpF0 ($X^2(1)= 14.257$;$p<0.001$) as well as logpF1 ($X^2(1)= 8.657$;$p=0.003$). There are no other significant interactions between ethnicity and any of the other predictors for this measure.

4.3.1.5.2.2. Sentence Data Linear Mixed Effects Analysis Results

There is no significant effect for ethnicity for the sentence data (X^2 (1)=0.023; p=0.880). Given the many significant effects for the interview data, it is perhaps unsurprising that a highly significant effect for the interaction between vowel and ethnicity was found for the sentence data (X^2 (2)=801.87; p<0.001) where vowel category was included as one of the predictors.

4.3.1.5.3. Summary of Findings for H2*−H4*

The Wilcoxon rank sum tests revealed a significant difference between black and white speakers for this measure, with black speakers having higher values than white speakers. This difference is also observable from a boxplot comparison of the values for the two groups, although the same difference is not clearly visible for the sentence data. There is however no significant effect for ethnicity for this variable according to the results of the linear mixed effects analysis, with the effect being a decrease in H2*−H4* values for white speakers. There are however significant

effects for fundamental frequency, first formant frequency, duration and speaker, while the effect for second formant frequency approaches significance. There are also significant interactions between ethnicity and both fundamental frequency as well as first formant frequency.

Garellek (2013:93), citing Zhang, Kreiman and Gerratt (2011), points out that this measure is considered to be correlated with the stiffness of the vocal folds as well as perceptual breathiness and Keating and Esposito (2006) report that, at least for certain languages, such as Bura for example, in conjunction with higher values for the measure H1−A3 (also found for black speakers in my study as presented later in this chapter), higher values for H2−H4 may signify breathier phonation. The overall pattern for both the uncorrected and corrected measures thus would suggest that black speakers may make use of breathier phonation as a predominant voice quality in comparison to white speakers and also that the stiffness of the vocal folds may be greater for white speakers, although an alternative interpretation is briefly discussed towards the end of this chapter. However, at least for the corrected measure, there are clearly more important predictors than that of ethnicity affecting H2*−H4* values.

4.3.1.5.4. Summary of Findings for H2–H4 (the uncorrected equivalent of H2*–H4*)

The Wilcoxon rank sum tests revealed a significant difference between black and white speakers for this measure, with black speakers having higher values than white speakers, a difference which is also visible from the boxplot comparison, although this difference is not as clear as for the corrected measure. There is a significant effect for ethnicity for this measure according to the results of the linear mixed effects analysis, with a decrease in H2−H4 values for white speakers. However, there are significant differences for all other predictors and there are also two significant interactions, namely between ethnicity and fundamental frequency as well as between ethnicity and second formant frequency.

4.3.2. Other Spectral Amplitude Measures

4.3.2.1. H1–A1* (the amplitude of the first harmonic minus the amplitude of the harmonic nearest F1, both corrected for the influence of formants and their bandwidths)*

4.3.2.1.1. H1*–A1* Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 4.34, illustrates the boxplots representing the values for the H1*−A1* sentence data, for all of the data measurement points grouped according to auditorily identified phonation type.

The phonation types which display the widest ranges of values (particularly in terms of the interquartile range) are those of breathy voice and whisper, suggesting more variability for these phonation types which involve aspiration noise primarily. Both breathy voice and whisper display a positive skew, suggesting a greater concentration of data points with lower values for this measure.

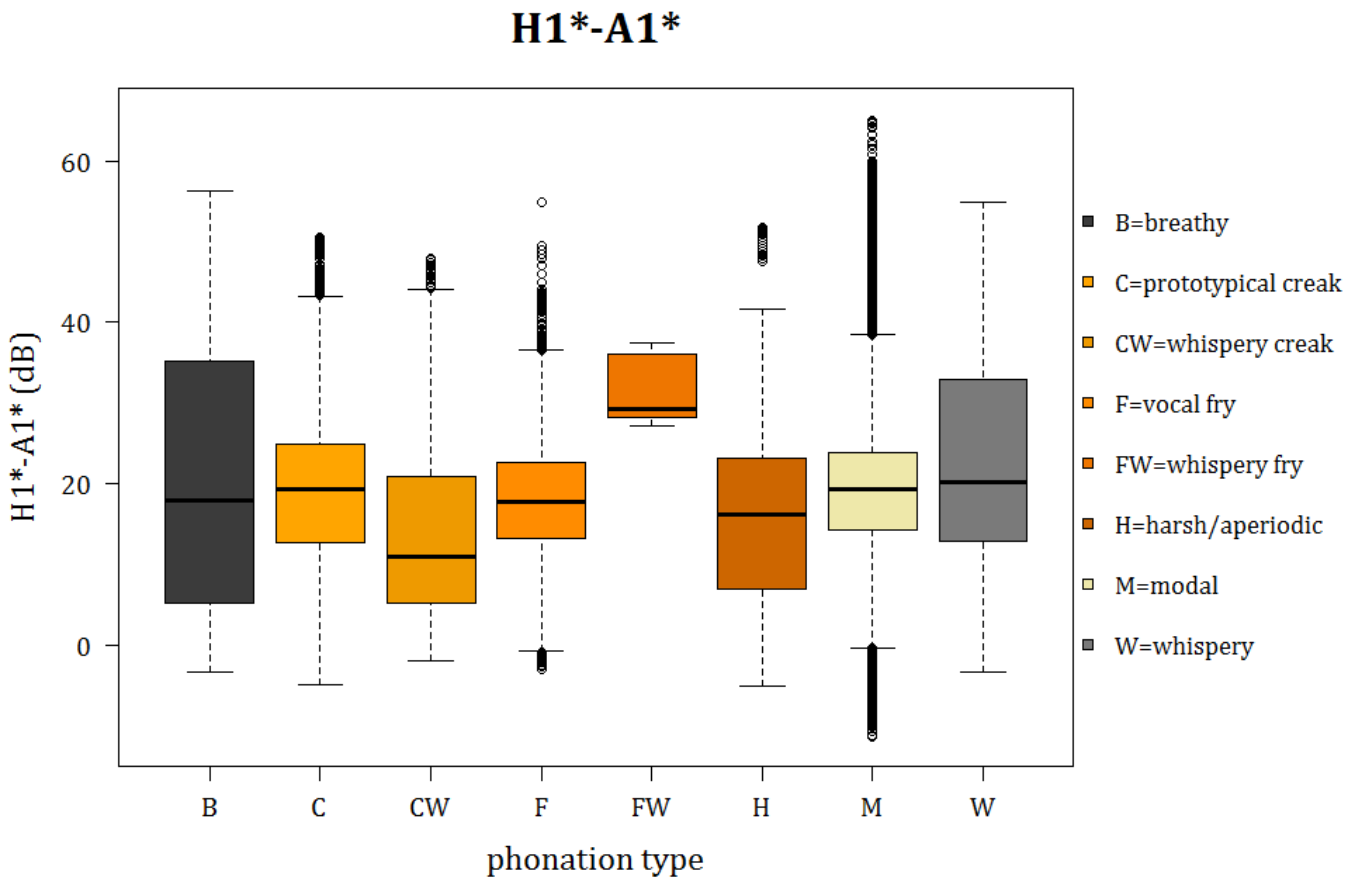

Figure 4.34: Boxplots displaying the values for the H1*−A1* sentence data for each of the auditorily identified phonation types for all data measurement points.

4.3.2.1.2. Statistical Analysis for H1*−A1*

The Wilcoxon rank sum test results reveal a difference between black and white speakers for this measure which approaches significance (W=209, *p*=0.071), for the hypothesis that the values for black speakers are greater than those for white speakers. Further evidence of this difference can be seen in the boxplot comparison of the values presented in figure 4.35 below.

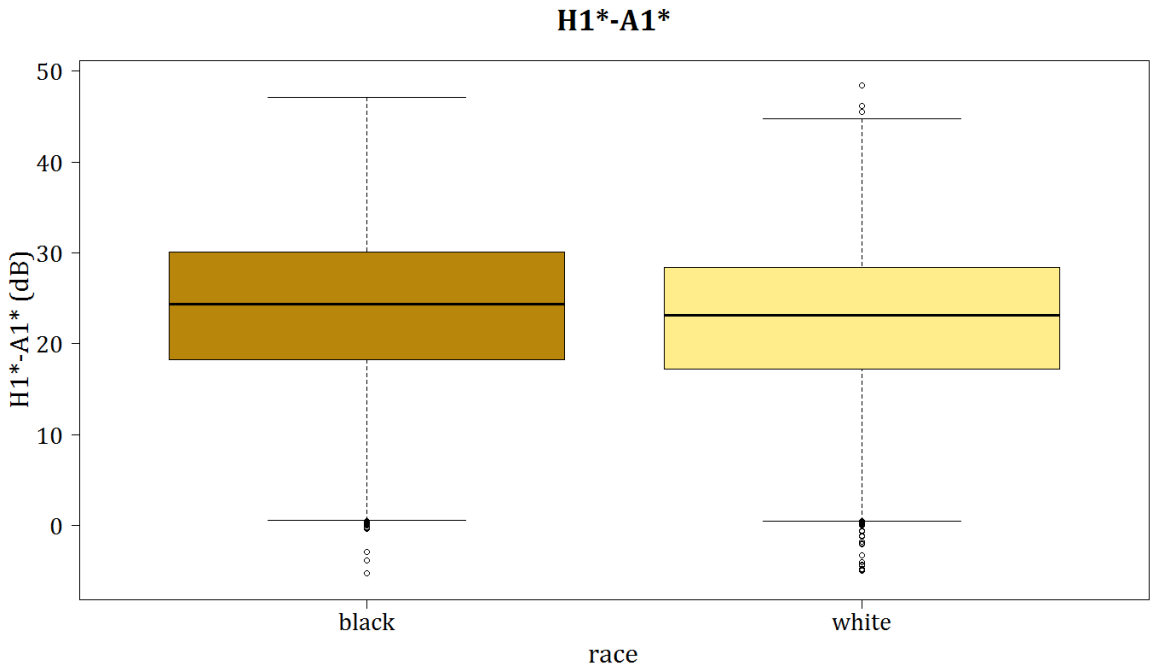

Figure 4.35: Boxplot comparison of the values for black and white speakers for the H1*−A1* interview data.

A similar pattern may be observed for the sentence data, as presented in figure 4.36 below, although the interquartile range for white speakers is wider.

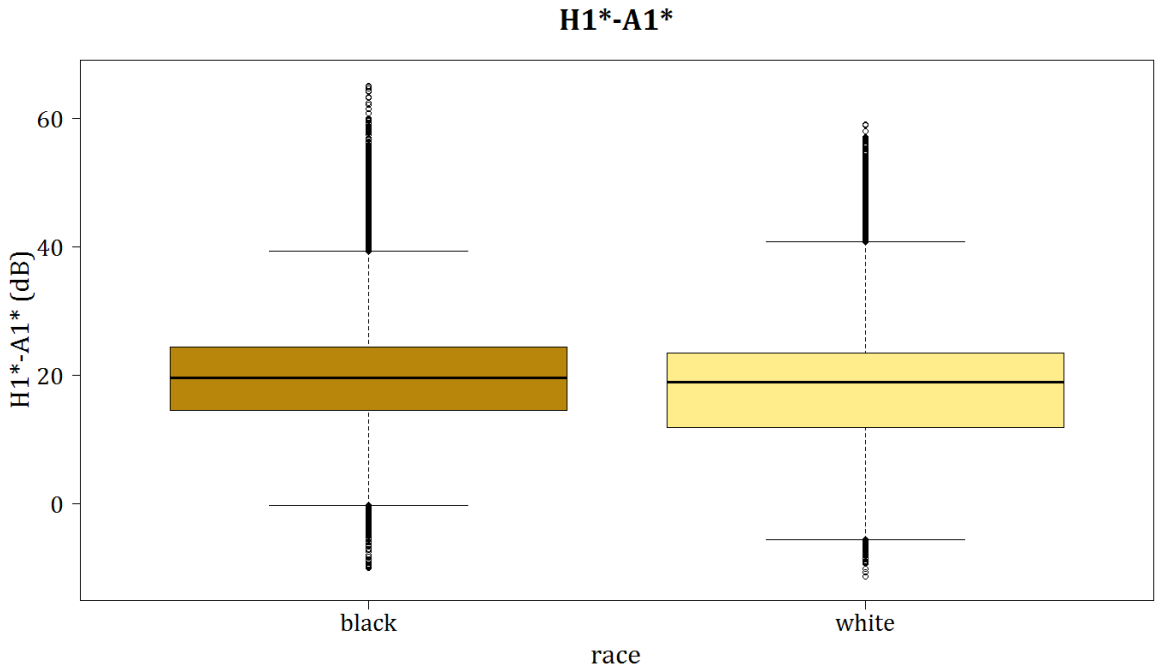

Figure 4.36: Boxplots representing the values for H1*−A1* sentence data for all data measurement points according to ethnicity.

4.3.2.1.2.1. Interview Data

The results of the linear mixed effects analysis reveals a significant effect for ethnicity ($X^2(1)= 5.2017; p=0.023$), such that there is a decrease of 1.796 dB ±0.75050 (standard errors) in the value for H1*−A1* for white speakers.

However, the results are significant for all other predictors for the interview data. The effect for logpF0 as a predictor is significant ($X^2(1)=1330.1; p<0.001$) with a 1% increase in pF0 increasing values by 0.06 dB ±0.15870 (standard errors), as are the effects for logEnergy ($X^2(1)=224.76; p<0.001$), a 1% increase in RMS Energy increasing values by 0.012 dB ±0.07645 (standard errors), logpF1 ($X^2(1)=4529; p<0.001$), a 1% increase in pF1 increasing values by 0.184 dB ±0.22382 (standard errors), logpF2 ($X^2(1)=242.79; p<0.001$) with a 1% increase in pF2 decreasing values by 0.039 dB ±0.24511 (standard errors), logduration ($X^2(1)=26.38; p<0.001$), a 1% increase in duration decreasing values by 0.008 dB ±0.14665 (standard errors) and speaker ($X^2(3)=858.75; p<0.001$).

There is also a significant interaction between ethnicity and logEnergy ($X^2(1)= 14.806; p<0.001$), logpF1 ($X^2(1)= 5.3254; p=0.021$) and logpF2 ($X^2(1)= 12.282; p<0.001$), while the interaction between ethnicity and logpF0 approaches significance ($X^2(1)= 3.3988; p=0.065$).

4.3.2.1.2.2. Sentence Data Linear Mixed Effects Analysis Results

For the sentence data, I found no significant effect for ethnicity (X^2 (1)=0, p=1), but a highly significant effect for the interaction between vowel category and ethnicity (X^2 (2)=1802.4, p<0.001).

4.3.2.1.2.3. Discussion

As can be seen from the scatterplots illustrated below in figures 4.37, 4.38 and 4.39, there is a clear positive correlation between F1 and this measure.

This is unsurprising given that Iseli, Shue and Alwan (2007:2291) point out that the measure H1*−A3* is dependent on F1. The correlation is similar for both groups of speakers.

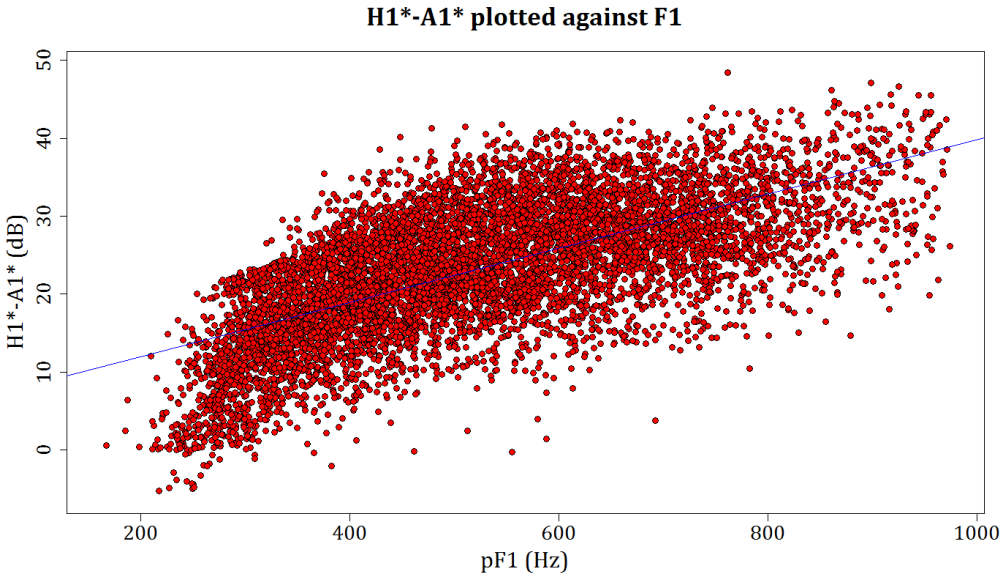

Figure 4.37: Scatterplot of H1*−A1* data for the whole sample plotted against pF1 in Hertz (Pearson's r= 0.663).

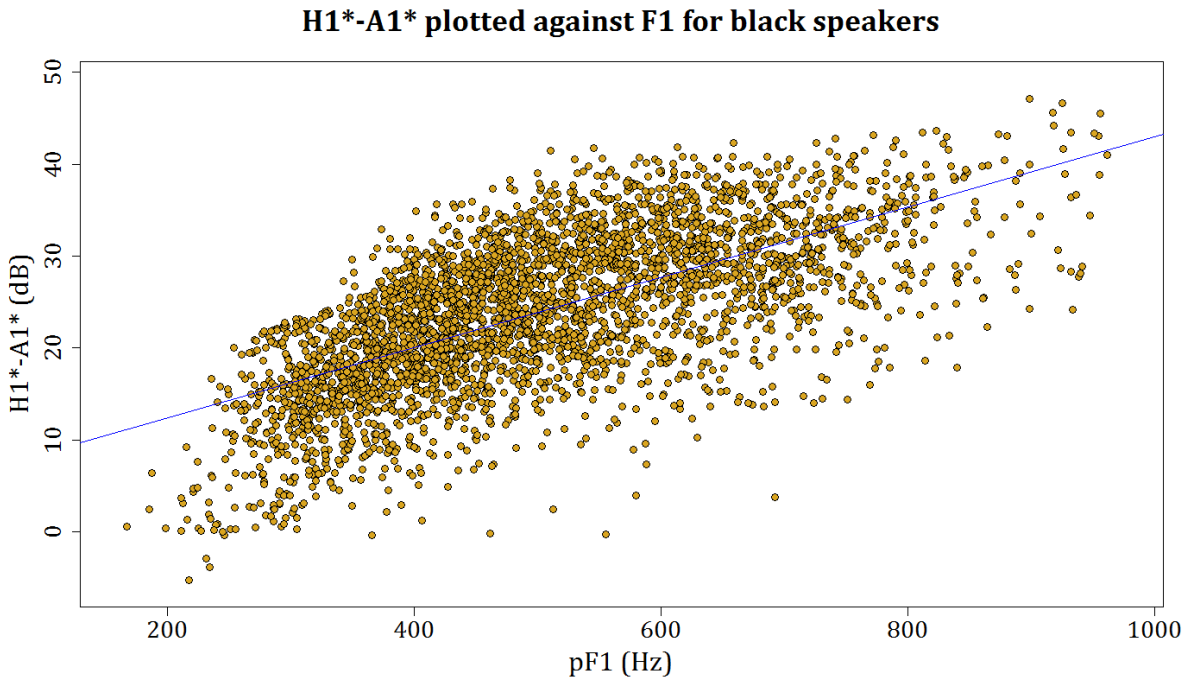

Figure 4.38: Scatterplot of H1*−A1* data for black speakers plotted against pF1 in Hertz (Pearson's *r*= 0.658).

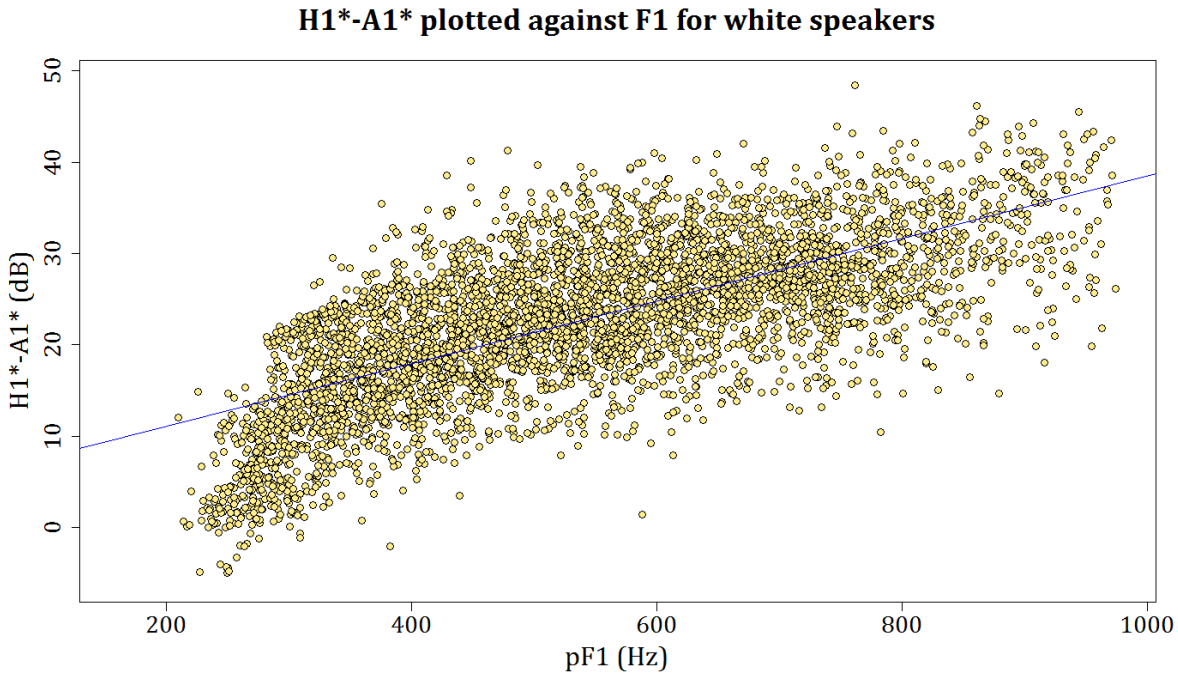

Figure 4.39: Scatterplot of H1*−A1* data for white speakers plotted against pF1 in Hertz (Pearson's *r*= 0.692).

When inspecting the values for H1*−A1* as a function of second formant frequency, as displayed below in figures 4.40, 4.41 and 4.42, it is clear that there is a moderate negative correlation for the sample as a whole as well as for the two ethnic groups.

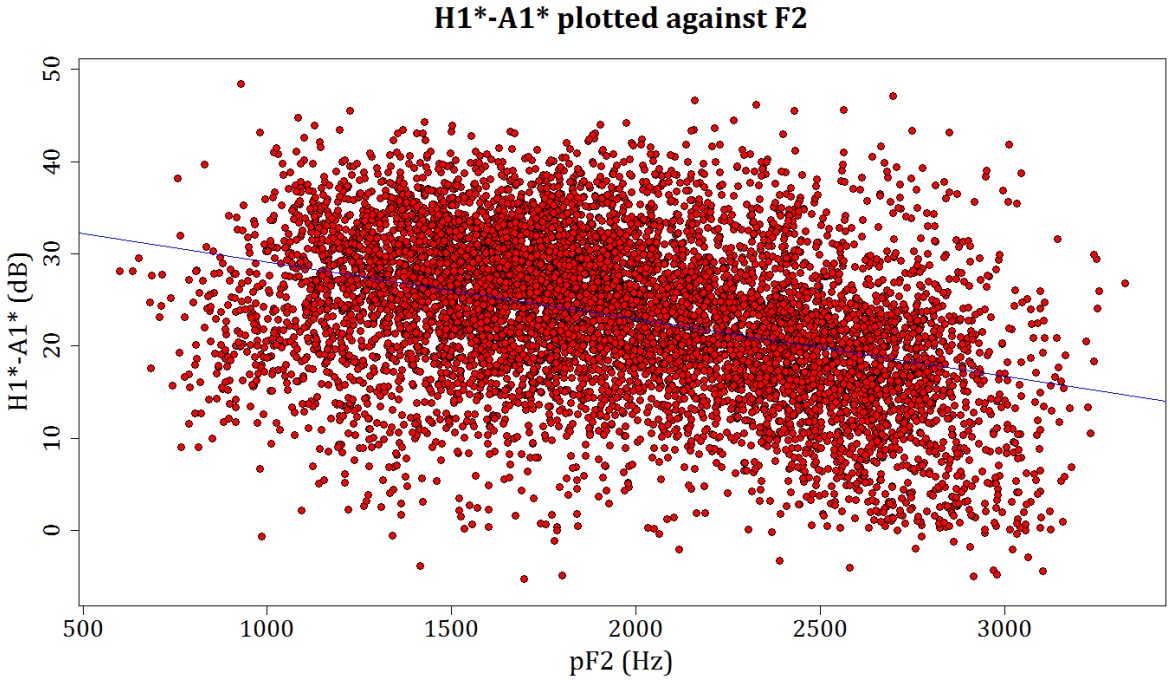

Figure 4.40: Scatterplot of H1*−A1* data for the whole sample plotted against pF2 in Hertz (Pearson's *r*= −0.378).

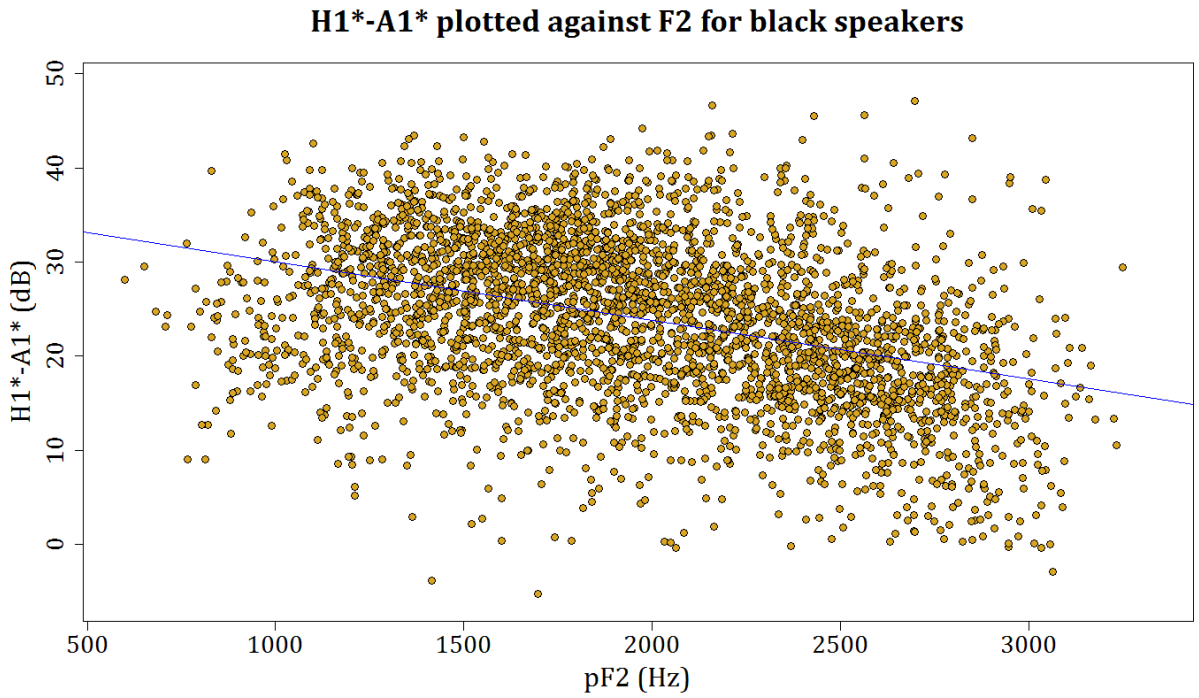

Figure 4.41: Scatterplot of H1*−A1* data for black speakers plotted against pF2 in Hertz (Pearson's *r*= −0.388).

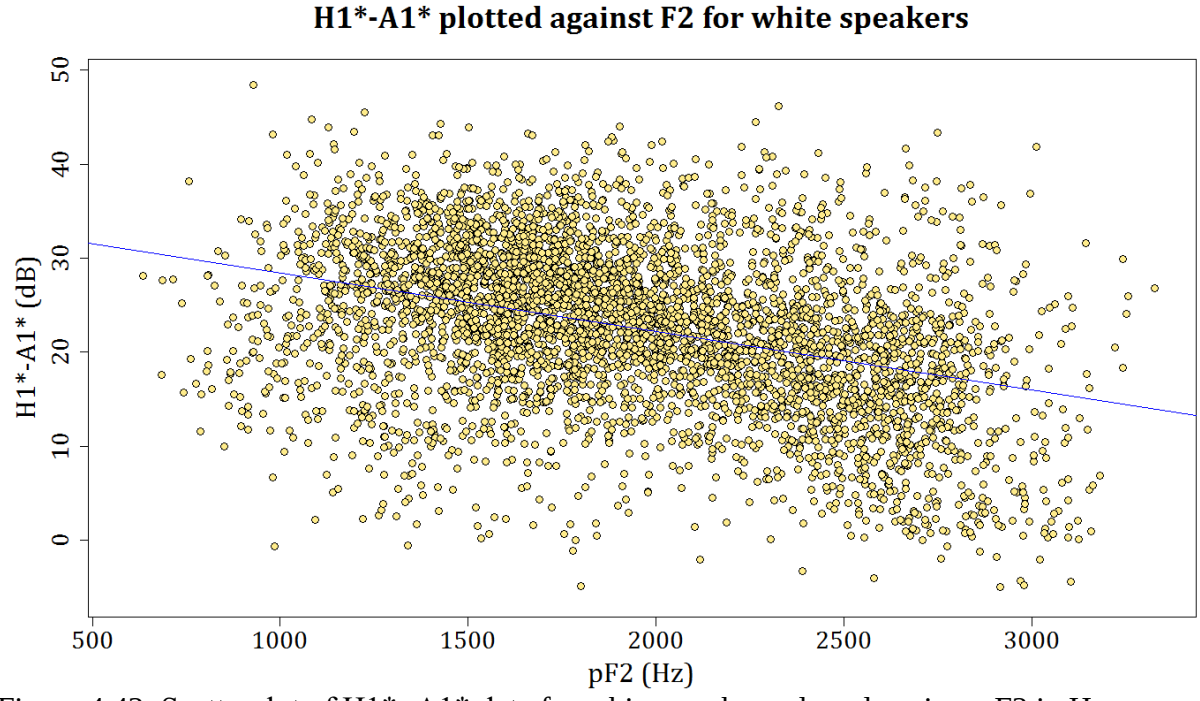

Figure 4.42: Scatterplot of H1*−A1* data for white speakers plotted against pF2 in Hertz (Pearson's *r*= −0.376).

4.3.2.1.3. Summary of Findings for H1*−A1*

The results of the Wilcoxon rank sum test for this measure reveal a difference approaching significance between black and white speakers, with black speakers having higher values than white speakers. This difference is also visible from the boxplot comparisons for both the interview data as well as the sentence data for H1*−A1*. The results of the linear mixed effects analysis agree with the Wilcoxon rank sum test results, in that there is a significant effect for ethnicity, with decreasing values for white speakers. There are however also significant effects for all other predictors. There are also significant interactions between ethnicity and several of the other predictors, including RMS energy, first formant frequency and second formant frequency. The interaction between ethnicity and fundamental frequency approaches significance.

4.3.2.1.4. Summary of Findings for H1−A1 (the uncorrected equivalent measure of H1*−A1*)

The results of the Wilcoxon rank sum tests also reveal a difference approaching significance between black and white speakers for this measure, with higher values for black speakers, a difference which is also evident from the boxplot comparison. While not significant, the effect for ethnicity approaches significance, with a decrease in H1−A1 values for white speakers. There are significant effects for all other predictors apart from second formant frequency, where the effect approaches significance. There are significant interactions between ethnicity and several of the other predictors, including energy, F1 and F2. The patterns for these interactions are generally very similar for both corrected and uncorrected measures.

According to Hanson et al (2001) as cited by Garellek (2012:7), the H1−A1 measure is expected to correlate with the presence of a glottal chink or posterior gap. Higher values for this measure are therefore expected for breathy voice due to the presence of greater glottal chinks in breathy voice (Garellek 2012:13). Thus the results for my study suggest that black speakers may have greater posterior glottal gaps in comparison to white speakers. This would agree with the results for some of the other measures included in this study, as discussed in the final chapter.

4.3.2.2. H1–A2* (the amplitude of the first harmonic minus the amplitude of the harmonic nearest F2, both corrected for the influence of formants and their bandwidths)*

4.3.2.2.1. H1*–A2* Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 4.43, displays the boxplots representing the values for the corrected harmonic differential measure H1*−A2* for all data measurement points according to auditorily identified phonation type for the sentence data. Differences between the auditorily identified phonation types are relatively slight, apart from a general tendency for phonation types hypothesized to involve more aspiration noise to have wider interquartile ranges.

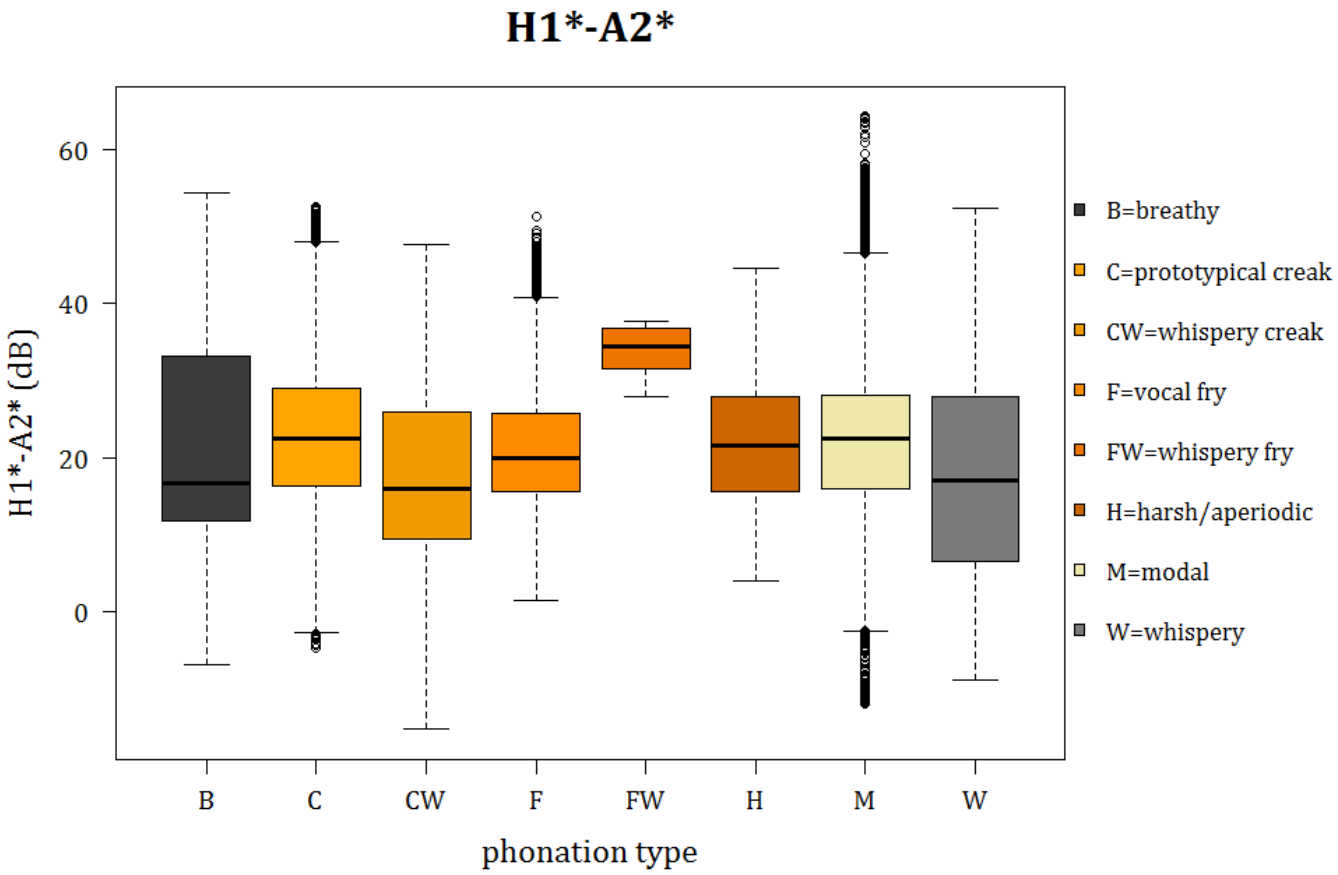

Figure 4.43: Boxplots displaying the values for the H1*−A2* sentence data for each of the auditorily identified phonation types for all data measurement points.

4.3.2.2.2. Statistical Analysis for H1*−A2*

The Wilcoxon rank sum test results reveal no significant effect (W=182, p=0.375) for ethnicity. However, based on the boxplot comparison provided in figure 4.44 below, there does appear to be an overall difference with black speakers exhibiting higher values.

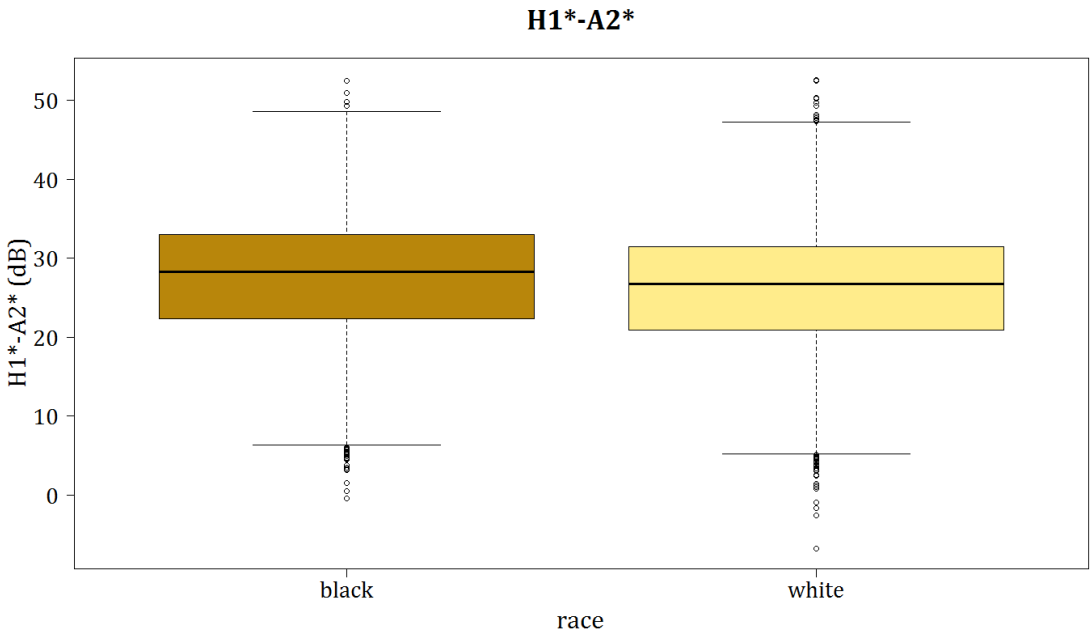

Figure 4.44: Boxplots representing the values for black and white speakers for H1*−A2* for the interview data.

This same pattern is also observed for the sentence data, as presented in figure 4.45 below.

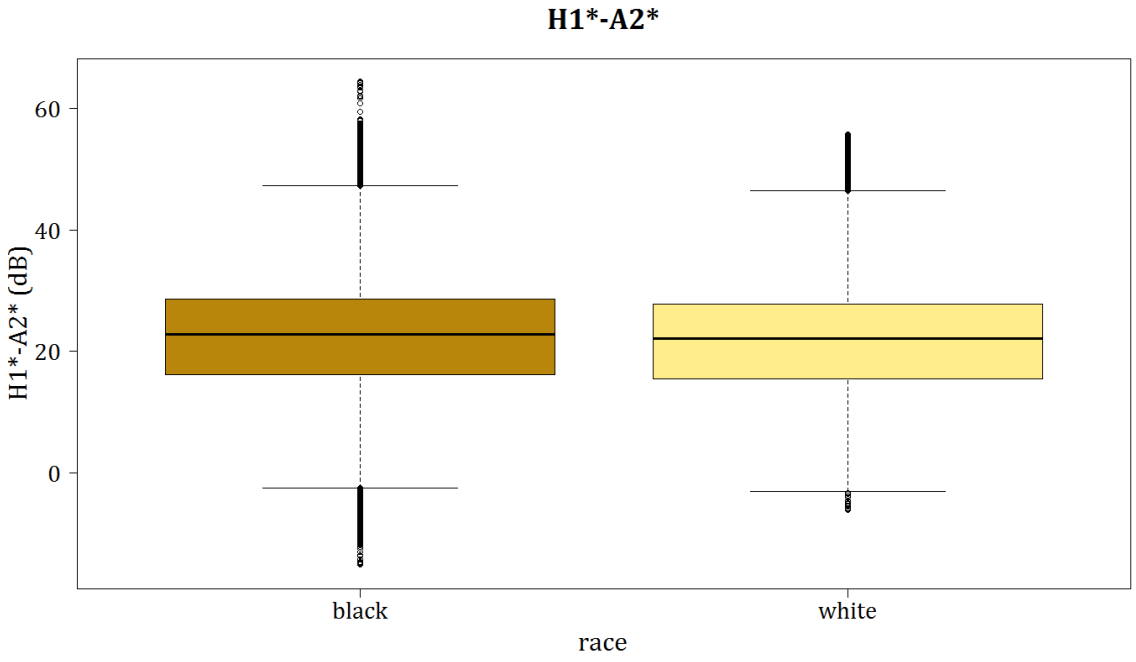

Figure 4.45: Boxplots representing the values for the H1*−A2* sentence data for all data measurement points according to ethnicity.

4.3.2.2.2.1. Interview Data

There is however no significant effect for ethnicity according to the results of the linear mixed effects regression analysis ($X^2(1)= 1.8333$;$p=0.176$) with a decrease of 1.211 dB ±0.87772 (standard errors) in H1*−A2* values for white speakers.

The effects of all other predictors are significant for the interview data. The effect for logpF0 is significant ($X^2(1)= 1458.3$;$p<0.001$), the effect of a 1% increase in pF0 being to increase H1*−A2* by 0.063 dB ±0.15812 (standard errors), as are the effects for logEnergy ($X^2(1)= 105.9$;$p=0.044$), a 1% increase in RMS Energy decreasing values by 0.008 dB ±0.07673 (standard errors), logpF1 ($X^2(1)= 3163.1$;$p<0.001$), a 1% increase in pF1 increasing H1*−A2* by 0.147 dB ±0.22915 (standard errors), logpF2 ($X^2(1)= 410.63$;$p<0.001$), a 1% increase in pF2 decreasing H1*−A2* by 0.053 dB ±0.25225 (standard errors), logduration ($X^2(1)= 58.551$;$p=0.001$), a 1% increase in duration decreasing H1*−A2* by 0.011 dB ±0.14722 (standard errors) and speaker ($X^2(3)= 1151$;$p<0.001$).

There are relatively few significant interactions between the other predictors and ethnicity, although one of the interactions approaches significance and another is significant. The interaction between ethnicity and logpF0 is significant ($X^2(1)= 10.76$;p=0.001) and the interaction between logpF1 and ethnicity is also significant ($X^2(1)= 8.4238$;p=0.004). The interaction between ethnicity and logEnergy approaches significance ($X^2(1)= 3.2521$;p=0.071).

4.3.2.2.2.2. Sentence Data Linear Mixed Effects Analysis Results

For the sentence data, as for the interview data, no significant effect was found for ethnicity (X^2 (1)=0.3054, p=0.581), although the effect for the interaction between ethnicity and vowel category was found to be highly significant (X^2 (2)=3059.8, p<0.001).

4.3.2.2.3. Summary of Findings for H1*−A2*

There is no significant difference between the values for black and white speakers for H1*−A2* according to the Wilcoxon rank sum tests. Black speakers do however exhibit higher values than white speakers for the boxplot comparisons. According to the linear mixed effects analysis, there is also no significant effect for ethnicity, with values of H1*−A2* decreasing for white speakers. There are significant effects for all other predictors. There are also significant interactions between ethnicity and fundamental frequency, as well as between ethnicity and first formant frequency. The interaction between ethnicity and RMS energy approaches significance.

Overall, while there is clearly a pattern for this measure which would suggest that black speakers use a voice quality involving more aspiration noise than white speakers, this effect does not reach statistical significance. The same applies for the uncorrected measure.

The measure H1−A2, included in this study in the form of H1*−A2* and its uncorrected equivalent H1−A2, is one of the earlier spectral measures to be used to capture perceptually relevant differences in phonation, as was done by Bickley (1982) who found that Gujarati listeners gave higher breathiness ratings for vowels with increased H1−A2 values. Based on Gobl and Ní Chasaide (1992), we would expect that for tense voice in comparison to modal voice, there should be smaller values for this measure, greater values for lax voice and that for both breathy and whispery voice, the values for this measure would be highest.

Blankenship (1997) used this measure as an indicator of the abruptness of vocal fold closure. Blankenship (1997) predicted that for certain phonation types such as that of breathy voice, where there is non-simultaneous closure of the vocal folds along their entire length, values for this measure should be higher, whereas in cases where there is greater vocal fold tension as is the case in laryngealisation (prototypical creak), the values for this measure could be expected to be lower and even negative rather than positive. Following Keating and Esposito (2006), this parameter is successful at distinguishing modal from breathy voice in several languages.

4.3.2.2.4. Summary of Findings for H1−A2 (the uncorrected equivalent of H1*−A2*)

The findings for the uncorrected measure overall are quite similar to those of the corrected measure. However, for the uncorrected measure there is a significant difference between black and white speakers according to the Wilcoxon rank sum test, with black speakers having higher values overall than white speakers. There is no significant effect for ethnicity according to the linear mixed effects analysis for the interview data, with decreasing values for white speakers. However, for the sentence data, the linear mixed effects analysis revealed a significant effect for ethnicity. There are significant effects for all other predictors. As for the corrected measure, there are also significant interactions between ethnicity and fundamental frequency as well as between ethnicity and first formant frequency. The interaction between ethnicity and energy approaches significance.

4.3.2.3. H1–A3* (the amplitude of the first harmonic minus the amplitude of the harmonic nearest F3, both corrected for the effects of formants and formant amplitudes)*

4.3.2.3.1. H1*–A3* Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 4.46, illustrates the boxplots representing the values for the H1*−A3* data for all data measurement points grouped according to auditorily identified phonation type. H1*−A3* is a measure of the skewness of the glottal pulse (Esling and Edmonson 2011) and following Blankenship (1997), there should be higher values for this measure for breathy vowels, lower values for modal voice and the lowest values for laryngealisation (i.e. creak). The figure demonstrates that, in general, the values for the different auditorily identified phonation types agree reasonably well with the predicted pattern based on the relevant literature for this measure.

Breathy voice exhibits one of the highest interquartile ranges apart from that of whispery fry. All of the distributions with the highest interquartile ranges are those types which are hypothesized to involve some degree of aspiration noise including whispery creak, whispery vocal fry, breathy voice and whisper itself. The lowest interquartile ranges are found for the non-composite creak types, such as prototypical creak as well as vocal fry. The interquartile range for modal voice is somewhere between these two extremes.

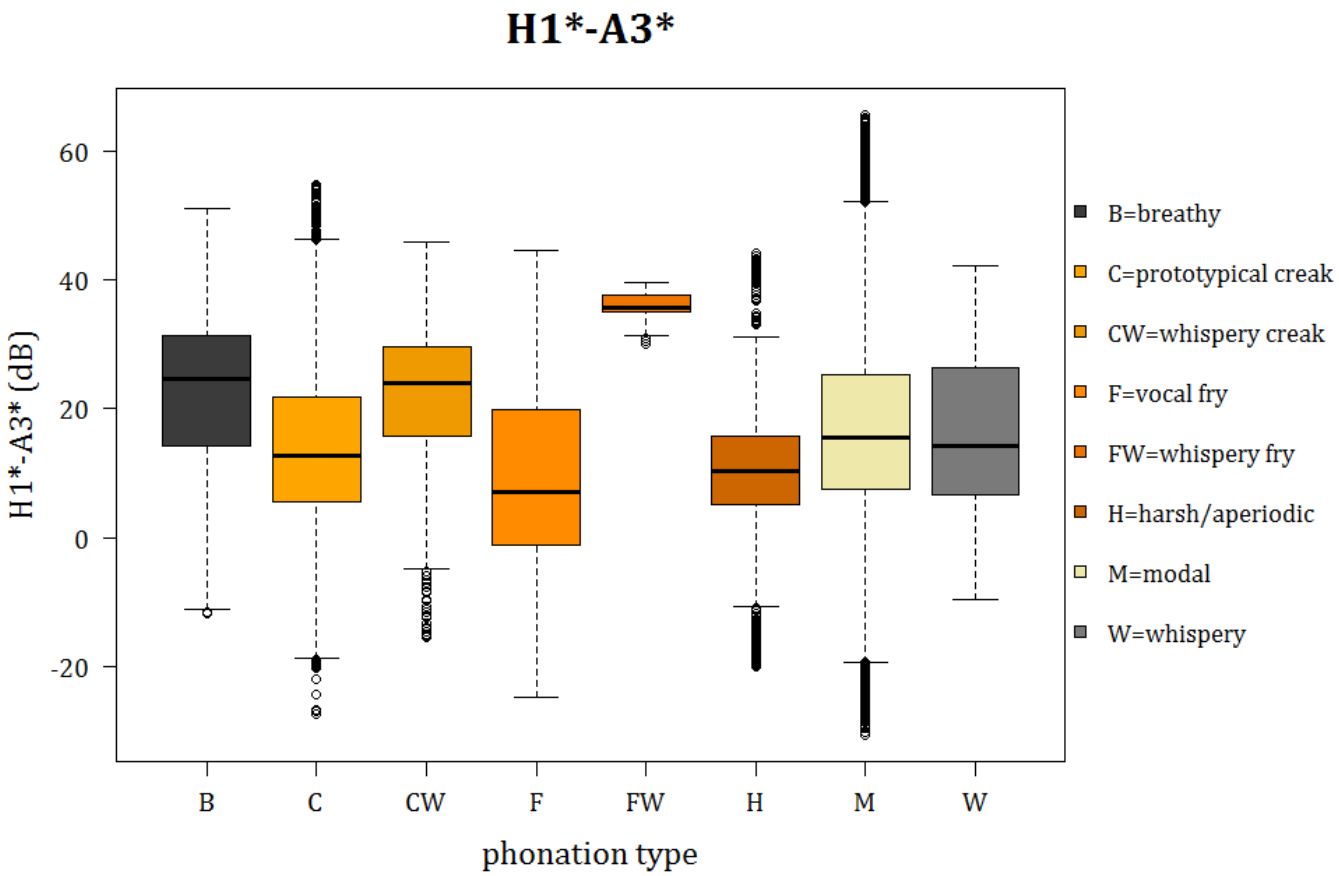

Figure 4.46: Boxplots displaying the values for the H1*−A3* sentence data for each of the auditorily identified phonation types for all data measurement points.

4.3.2.3.2. Statistical Analysis for H1*−A3*

The Wilcoxon rank sum test reveals no significant effect for ethnicity for this measure (W=188, *p*=0.212). However, black speakers exhibit higher values than white speakers overall which can be observed from the boxplot comparison provided in figure 4.47 below.

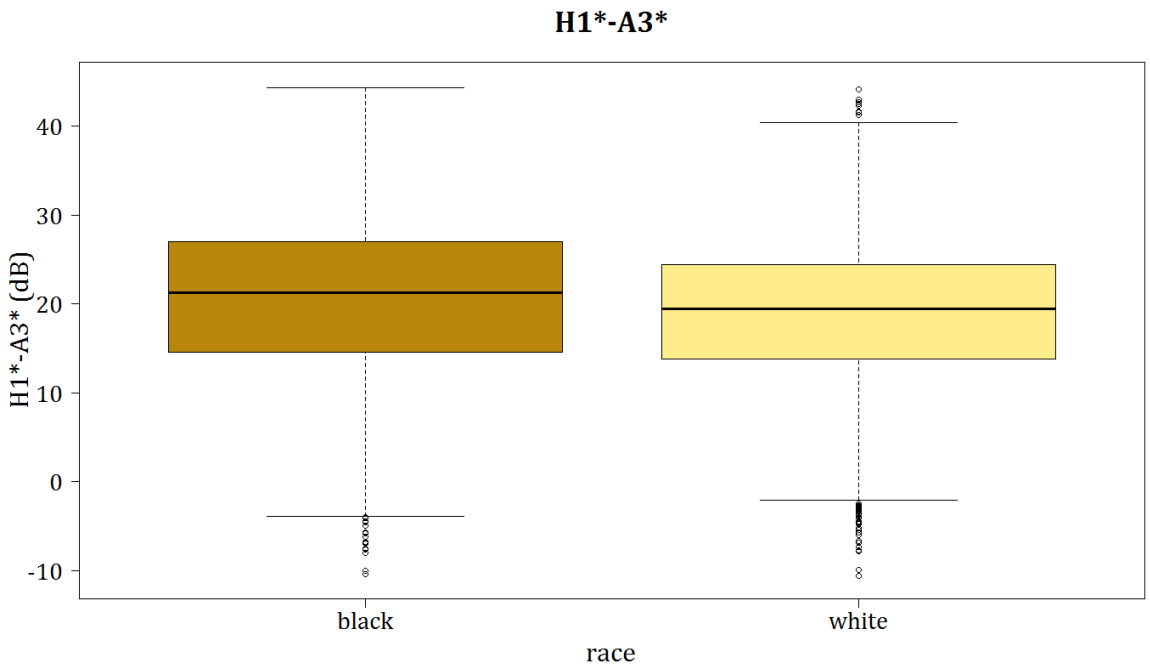

Figure 4.47: Boxplots representing the values for black and white speakers for the H1*−A3* interview data.

This difference can also be observed for the sentence data as displayed in figure 4.48 below, although it is not as clear.

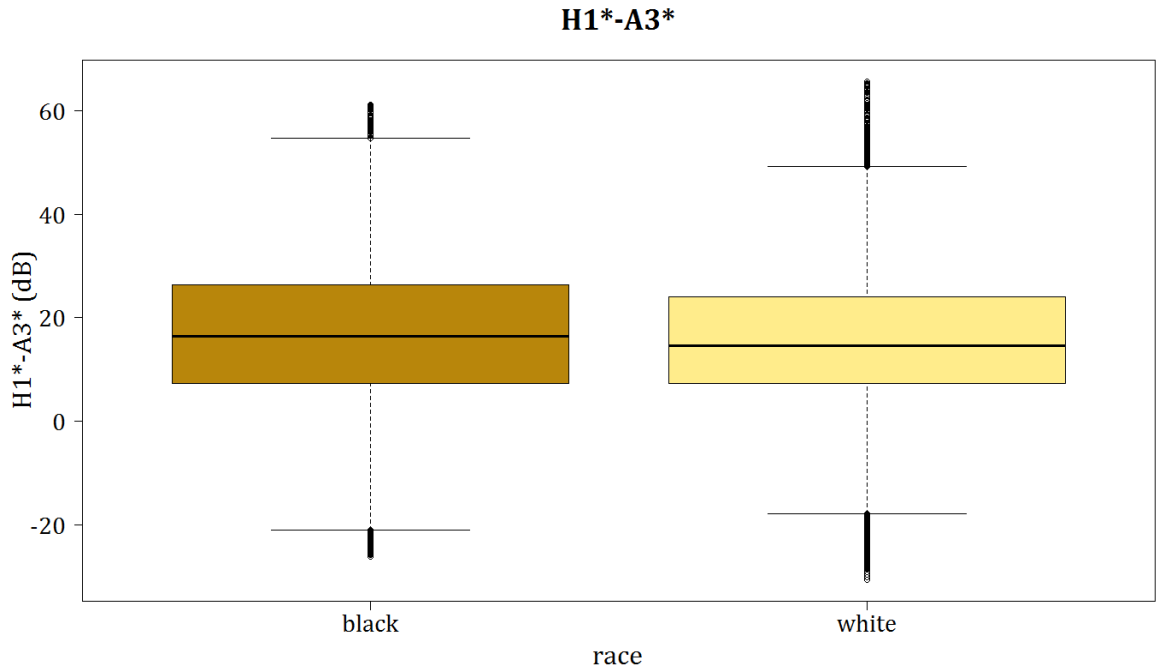

Figure 4.48: Boxplots representing the values for H1*−A3* sentence data for all data measurement points according to ethnicity.

4.3.2.3.2.1. Interview Data

According to the linear mixed effects analysis however, there is no significant effect for ethnicity ($X^2(1)= 0.0776; p=0.781$) with a decrease of 0.263 dB ±0.94111 (standard errors) in H1*−A3* values for white speakers.

There is a significant effect for logpF0 ($X^2(1)= 957.94; p<0.001$), a 1% increase in pF0 increasing H1*−A3* by 0.056 dB ±0.17574 (standard errors), as well as for logpF1 ($X^2(1)= 2515.2; p<0.001$), a 1% increase in pF1 increasing H1*−A3* by 0.141 dB ±0.25315 (standard errors), logpF2 ($X^2(1)= 2347.1; p<0.001$) , a 1% increase in pF2 increasing H1*−A3* values by 0.153 dB ±0.27835 (standard errors), logduration ($X^2(1)= 69.694; p<0.001$), with a 1% increase in duration decreasing H1*−A3* by 0.014 dB ±0.16339 (standard errors) and speaker ($X^2(3)= 2397.5; p<0.001$) for the interview data. Only the effect for logEnergy is not significant ($X^2(1)= 0.0472; p=0.828$).

There are several significant interactions between ethnicity and the other predictors for this variable. There is a highly significant interaction between logpF2 and ethnicity ($X^2(1)= 27.982; p<0.001$) and a highly significant interaction between ethnicity and logpF0 ($X^2(1)= 17.73; p<0.001$). There are significant interactions between ethnicity and logduration ($X^2(1)= 10.619; p=0.001$) as well as between ethnicity and logpF1 ($X^2(1)= 3.8751; p=0.049$).

4.3.2.3.2.2. Sentence Data Linear Mixed Effects Analysis Results

There is no significant effect for ethnicity for the sentence data (X^2 (1)=0.8651, p=0.352), although there are highly significant effects for the interaction between vowel category and ethnicity (X^2 (2)=3660.9, p<0.001).

4.3.2.3.3. Summary of Findings for H1*−A3*

The results of the Wilcoxon rank sum tests reveal no significant difference between black and white speakers for this measure. The boxplot comparisons however reveal that black speakers tend to have higher values than white speakers overall. According to the results of the linear mixed effects analysis however, there is no significant effect for ethnicity, with decreasing values of H1*−A3* for white speakers. There are significant effects for all other predictors, apart from for RMS energy. There are several significant interactions between ethnicity and the other predictors, including F1, F2, duration and f_0.

The overall pattern is that white speakers display lower values for this measure in comparison to black speakers and thus can be hypothesized to have more abrupt glottal closures than black speakers. The higher values for H1*−A3* can plausibly be linked to non-simultaneous closure of the ligamental glottis (Hanson 1997:479).

Following Hanson and Chuang (1999:1067), speakers with lower values for H1*−A3* and H1*−A1 were placed into 'group 1' and these speakers were also found to have lower noise ratings, suggesting less aspiration noise. Thus Hanson and Chuang (1999:1067) hypothesize that such speakers have a relatively abrupt closure of the glottis with posterior glottal openings varying in size. These relative differences resemble the values found for the group of white speakers in my study. The black speakers on the whole, resemble Hanson and Chuang's (1999:1067) group two speakers, in terms of having relatively high values (by comparison at least) for H1*−A3* and H1*−A1* (in Hanson and Chuang's 1999:1067 study, A1 was not corrected for) and while perceptual ratings for noise as such were not included in my study, black speakers do show a pattern for all of the noise measures included in this study (SHR, CPP and HNR, as presented in the following chapter) which would suggest that their speech is characterized by a more prominent noise component than their white counterparts.

Speakers with both relatively low H1*−A3* values and low H1*−A1 values, according to Hanson (Hanson 1997:478) can be said to have relatively flat spectral tilts as well as prominent peaks for the first formant. Such speakers can thus be hypothesized to have abrupt closure of the glottis (Hanson 1997:478). This is the overall pattern found for white speakers, suggesting that in comparison to black speakers, the predominant voice quality for white speakers involves more abrupt glottal closure.

4.3.2.3.4. Summary of Findings for H1−A3 (the uncorrected equivalent of H1*−A3*)

The results of the Wilcoxon rank sum tests reveal no significant differences although black speakers appear to have higher values overall. This difference is evident from the boxplot comparison, although it is less obvious than for the corrected measure and the pattern is different for the sentence data. There is no significant effect for ethnicity based on the results of the linear mixed effects analysis with a decrease in H1−A3 values for white speakers. There are significant

effects for all other predictors apart from duration. There are significant interactions between ethnicity and both F1 and RMS energy. The results for the uncorrected measure are therefore similar to those for the corrected measure.

4.4. CONCLUSION

In this chapter, I have presented the results of the auditory analysis of the sentence data and have also provided the results of the acoustic analysis, focusing on the interview data for the harmonic differential measures. In the following chapter, I will present the results of the measures of noise based primarily on the acoustic results from the interview data and will provide an overview of the research findings.

CHAPTER V: RESULTS OF THE ACOUSTIC ANALYSIS FOR THE NOISE MEASURES AND AN OVERVIEW OF THE RESEARCH FINDINGS

5.1. INTRODUCTION

In this chapter, I present the findings for the noise measures as used in this study, namely SHR, CPP and HNR as described in chapter two. In presenting these results, I follow the same format as that of the previous chapter detailing the results of the harmonic differential measures. I conclude with a summary of the results from both the present chapter and the previous chapter and suggest possible initial interpretations for the research findings.

5.2. SHR (SUBHARMONICS-TO-HARMONICS RATIO)

5.2.1. SHR Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 5.1, displays the boxplots representing the values for the auditorily identified phonation types for the SHR sentence data. The total ranges for the different phonation types are very similar to one another in most cases, although the highest values overall are those of modal voice with the lowest for harsh/aperiodic voice. Breathy voice and whisper as well as most creak types are in between these two extremes.

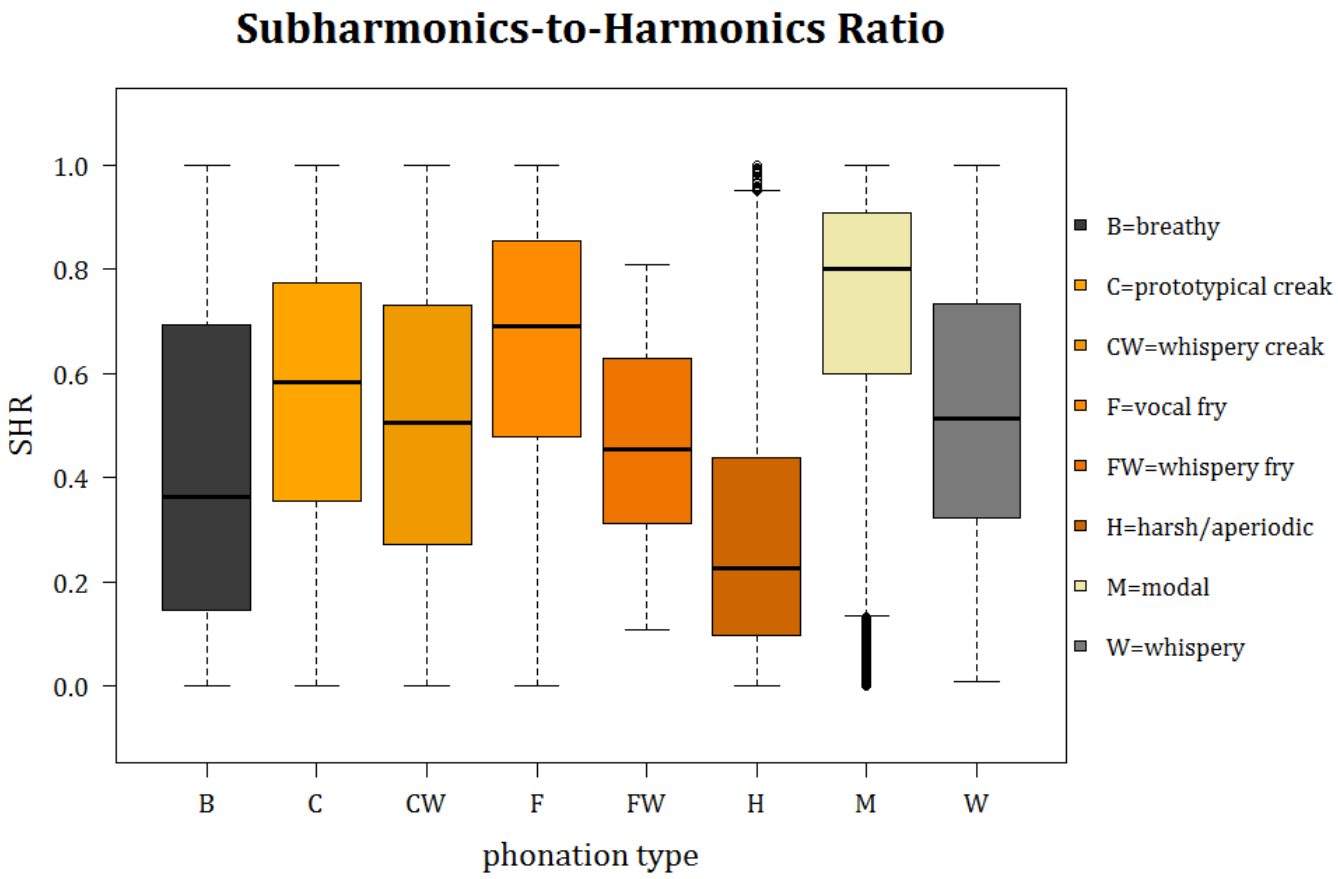

Figure 5.1: Boxplots displaying the values for the SHR sentence data for each of the auditorily identified phonation types for all data measurement points.

5.2.2. Statistical Analysis for SHR

The results of the Wilcoxon rank sum test reveal that there is no significant difference between black and white speakers for this measure (W=127, p=0.139). However white speakers have higher values overall, as can be seen from the boxplot comparison presented in figure 5.2 below. These results would suggest that there are a greater number of subharmonics to harmonics for white speakers.

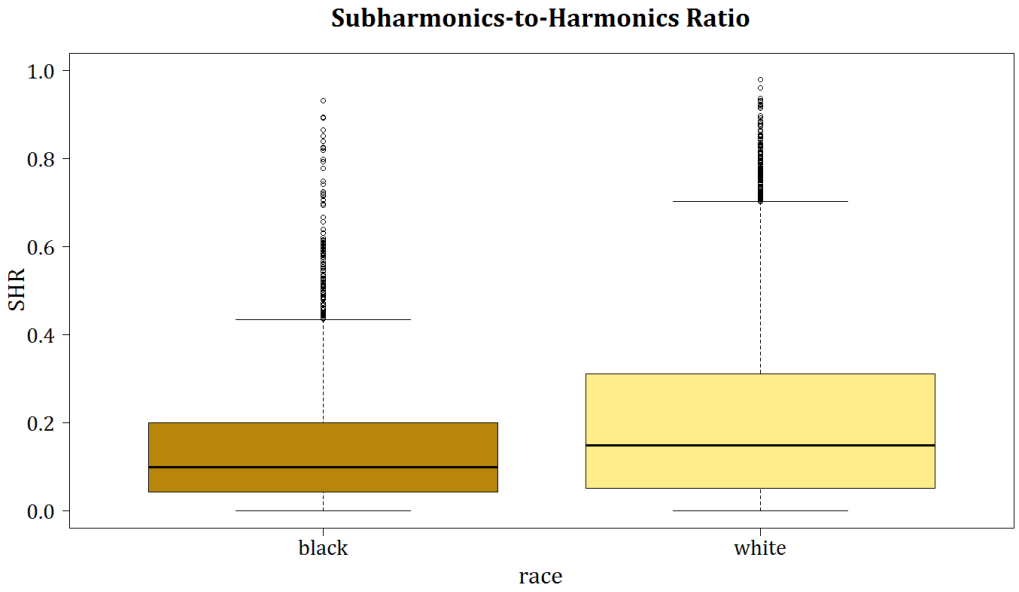

Figure 5.2: Boxplots representing the values for black and white speakers for the SHR interview data.

This pattern appears to be different for the sentence data values displayed in figure 5.3 below. That there is an apparent difference between the patterning of the interview data and sentence data could potentially be attributed to differences in terms of multiple-pulsing between more formal speech and casual speech (and the fact that no IP final tokens were included in the sentence data). However there are fewer data measurement points for SHR compared to the other measures due to the level of sensitivity of the SHR algorithm, so this discrepancy ought not to be assigned too much weight.

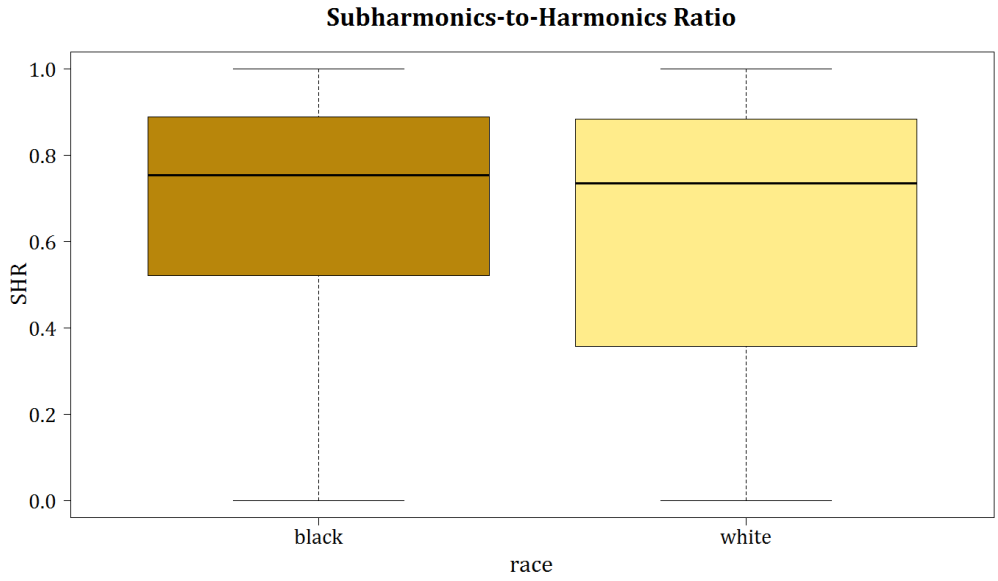

Figure 5.3: Boxplots representing the values for SHR sentence data for all data measurement points according to ethnicity.

5.2.2.1. Interview Data

The results of the linear mixed effects analysis reveal that while there is no significant effect for this measure, the effect does approach statistical significance ($X^2(1)= 2.9912$;$p=0.084$), with an increase of 0.037 ±0.020400 (standard errors) in the values for SHR for white speakers.

There are also significant effects for some of the other predictors. There is a significant effect for logEnergy ($X^2(1)= 183.69$;$p<0.001$), a 1% increase in RMS Energy increasing SHR by 0.0004 ±0.003014 (standard errors), logduration ($X^2(1)= 124.59$;$p<0.001$), a 1% increase in duration decreasing SHR by 0.0006 ±0.005777 (standard errors) and speaker ($X^2(3)= 560.4$;$p<0.001$). No significant effects were found for logpF2 ($X^2(1)= 1.8156$;$p=0.178$), logpF1 ($X^2(1)= 2.3556$;$p=0.125$) and logpF0 ($X^2(1)= 1.585$;$p=0.208$).

There is also a significant interaction between ethnicity and logpF0 ($X^2(1)= 6.207$;$p=0.013$) and a highly significant interaction between ethnicity and logEnergy ($X^2(1)= 21.934$;$p<0.001$). There are no other significant interactions between ethnicity and the other predictors for this measure.

5.2.2.2. Sentence Data Linear Mixed Effects Analysis Results

For the sentence data, the result of the linear mixed effects analysis is not significant (X^2 (1)=0.1127, p=0.737). There is also no significant interaction between ethnicity and vowel category for this measure (X^2 (2)=0.686, p=0.71).

5.2.3. Summary of Findings for SHR

The Wilcoxon rank sum tests revealed no significant difference between black and white speakers for this measure, with white speakers exhibiting higher values overall than black speakers, as is evident from a boxplot comparison. This difference is however not as clear for the sentence data. The effect of ethnicity approaches significance for this measure, according to the results of the linear mixed effects analysis, with increasing SHR values for white speakers. There are also significant effects for RMS energy, duration as well as speaker. There are significant interactions between ethnicity and fundamental frequency as well as between ethnicity and RMS energy.

5.3. CPP (CEPSTRAL PEAK PROMINENCE)

5.3.1. CPP Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 5.4 displays the boxplots representing the values for CPP across auditorily identified phonation types for all of the data measurement points. This measure appears to be fairly useful in terms of distinguishing between different phonation types as identified by means of the auditory coding procedure employed. Modal voice displays the widest range of values, both for the interquartile range and the overall range. Lower interquartile ranges can be found for non-compound creak types such as prototypical creak and vocal fry. The phonation types with the lowest interquartile ranges of all are those which are hypothesized to involve a high degree of aspiration noise, namely breathy voice and whisper. For these two phonation types, the boxplots are relatively similar, with breathy voice displaying a somewhat wider range.

Phonation types hypothesized to involve a mixture of creak and aspiration noise components unsurprisingly display interquartile ranges which are between those of breathy voice and whisper on the one hand and modal voice on the other. Harsh/aperiodic voice presents with

relatively low values and in particular, a low interquartile range. In agreement with Keating et al. (2015), prototypical creak shows lower CPP values than those for modal voice, which can presumably be accounted for in terms of the involvement of jitter in the production of prototypical creak. This would also help explain why the values for vocal fry are slightly higher than those of creak, given that vocal fry is hypothesized to exhibit a greater regularity in fundamental frequency. Whispery creak is naturally lower, as expected, given the addition of aspiration noise to a phonation type involving irregular f_0.

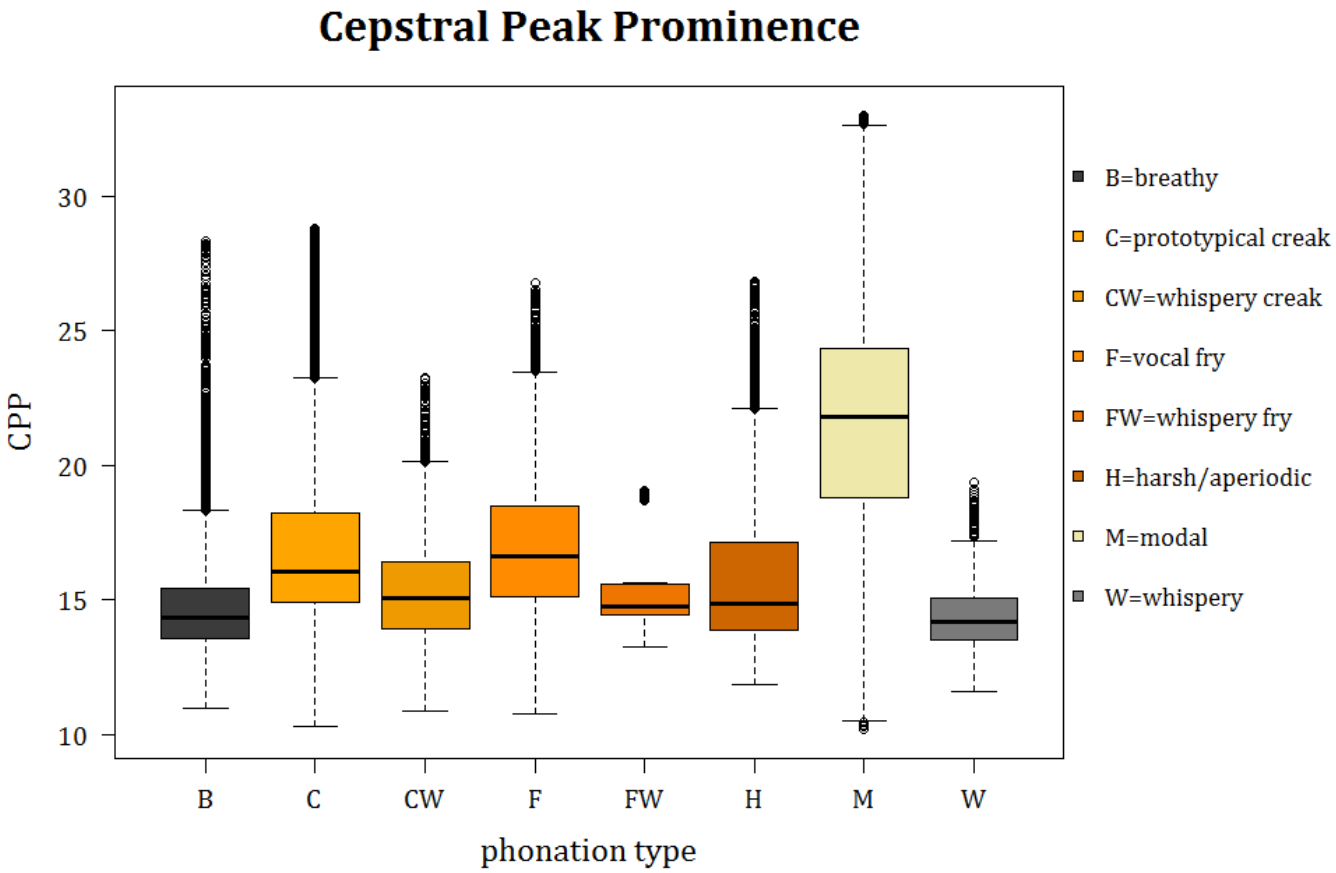

Figure 5.4: Boxplots displaying the values for the CPP sentence data for each of the auditorily identified phonation types for all data measurement points.

5.3.2. Statistical Analysis for the CPP Data

The Wilcoxon rank sum test revealed a significant difference between black and white speakers for this measure (W=84, p=0.006), for the alternative hypothesis that the values for white speakers are higher than those for black speakers. This difference can also be observed from the

boxplot comparison presented in figure 5.5 below.

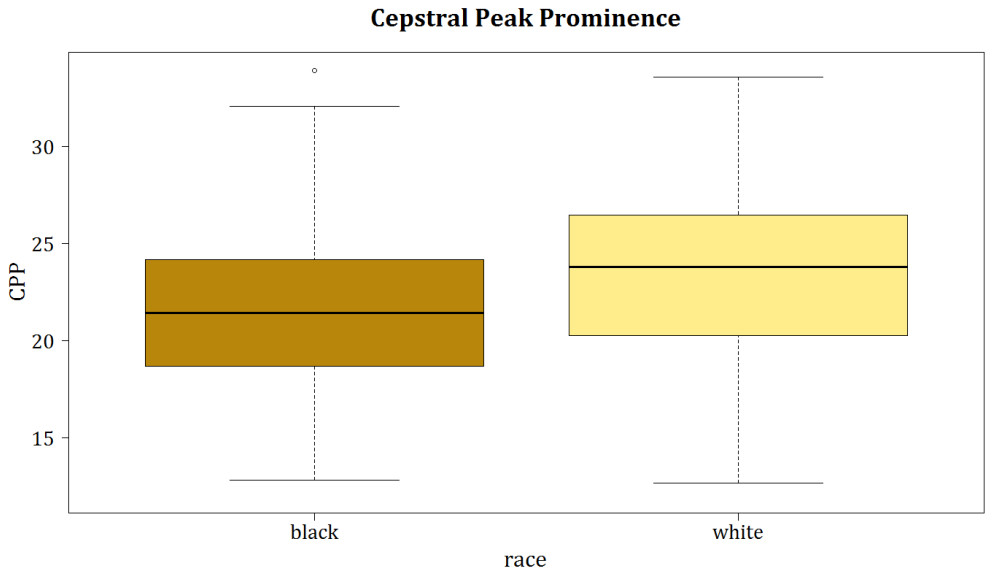

Figure 5.5: Boxplots representing the values for black and whites speakers for the CPP interview data.

This pattern is also observable for the CPP sentence data, as displayed in figure 5.6 below.

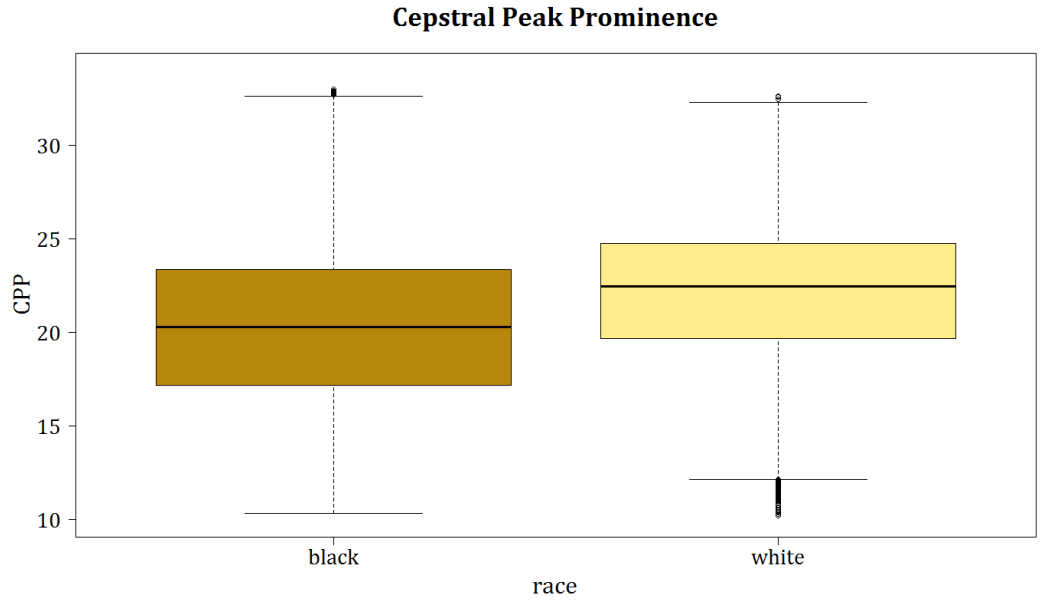

Figure 5.6: Boxplots representing the values for CPP sentence data for all data measurement points according to ethnicity.

5.3.2.1. Interview Data

There is a significant effect for ethnicity for this measure based on the linear mixed effects analysis ($X^2(1)= 4.7167$;$p=0.03$) of the interview data, with an increase of 1.66 ±0.73787 (standard errors) in CPP values for white speakers.

For most of the other predictors for the interview data, there are also significant effects, apart from logduration ($X^2(1)= 1.2176$;$p=0.27$). There is a significant effect for logpF0 ($X^2(1)= 45.955$;$p<0.001$), a 1% increase in pF0 increasing CPP values by 0.006 ±0.09178 (standard errors), logEnergy ($X^2(1)= 1677$;$p<0.001$), a 1% increase in RMS Energy increasing values by 0.018 ±0.04105 (standard errors), logpF1 ($X^2(1)= 327.96$;$p<0.001$), a 1% increase in pF1 increasing CPP values by 0.022 ±0.12075 (standard errors), logpF2 ($X^2(1)= 15.205$;$p<0.001$), a 1% increase in pF2 decreasing CPP values by 0.005 ±0.13199 (standard errors) and speaker ($X^2(3)= 3739.3$;$p<0.001$).

There are also some significant interactions with ethnicity. There is a significant interaction between ethnicity and logpF0 ($X^2(1)= 18.169$;$p<0.001$), logpF1 ($X^2(1)= 5.0344$;$p=0.025$), logpF2 ($X^2(1)= 6.0417$;$p=0.014$), logEnergy ($X^2(1)= 16.872$;$p<0.001$) and duration ($X^2(1)= 15.2$;$p<0.001$).

5.3.2.2. Sentence Data Linear Mixed Effects Analysis Results

This significant effect found for ethnicity for the interview data does not apply to the sentence data (X^2 (1)=0.08208, $p=0.266$). However, for the sentence data, the interaction between ethnicity and vowel was found to be highly significant (X^2 (2)=1744.1, $p<0.001$).

5.3.2.3. Discussion

Turning to the significant interactions between ethnicity and the other predictors for CPP, as can be observed in the scatterplots provided in appendix E, there is a weak positive correlation between CPP and fundamental frequency for the sample as a whole. This is in line with Keating, Garellek and Kreiman's (2015) findings using resynthesis suggesting that lowering fundamental frequency may result in a lowering of CPP. However, the correlations are in opposite directions for black and white speakers and in some cases (particularly for black speakers), there appears to be some bimodality present, making this relatively weak correlation difficult to straightforwardly interpret.

For black speakers there is a weak positive correlation between CPP and pF0, while for white speakers, there is a weak negative correlation. It is not entirely clear what phonatory behaviour is at play to lead to such a difference, although it may presumably be linked to tonal effects in African languages. However, black speakers overall have a lower pitch than white speakers. Thus one may expect lower CPP values overall for black speakers and in fact, this is what is observed. This may suggest that black speakers, to a greater extent than white speakers, use non-modal phonation types involving noise, at lower pitches, such that to increase pitch would mean to become more modal. As Awan, Giovinco and Owens (2012:670.e18-670.e19) point out, increasing f_0 should increase values for CPP due to there being greater stability for f_0 when increased since there are increased motor-unit firing rates and a resulting decrease in jitter. In this regard, the data for the black speakers can be accounted for.

As can be seen from figures 5.7, 5.8 and 5.9 below, displaying the scatterplots for the CPP data when plotted against pF2, there is a negative correlation between second formant frequency and CPP. This applies to both black and white speakers, but the values are lower with a slightly stronger correlation overall for black speakers.

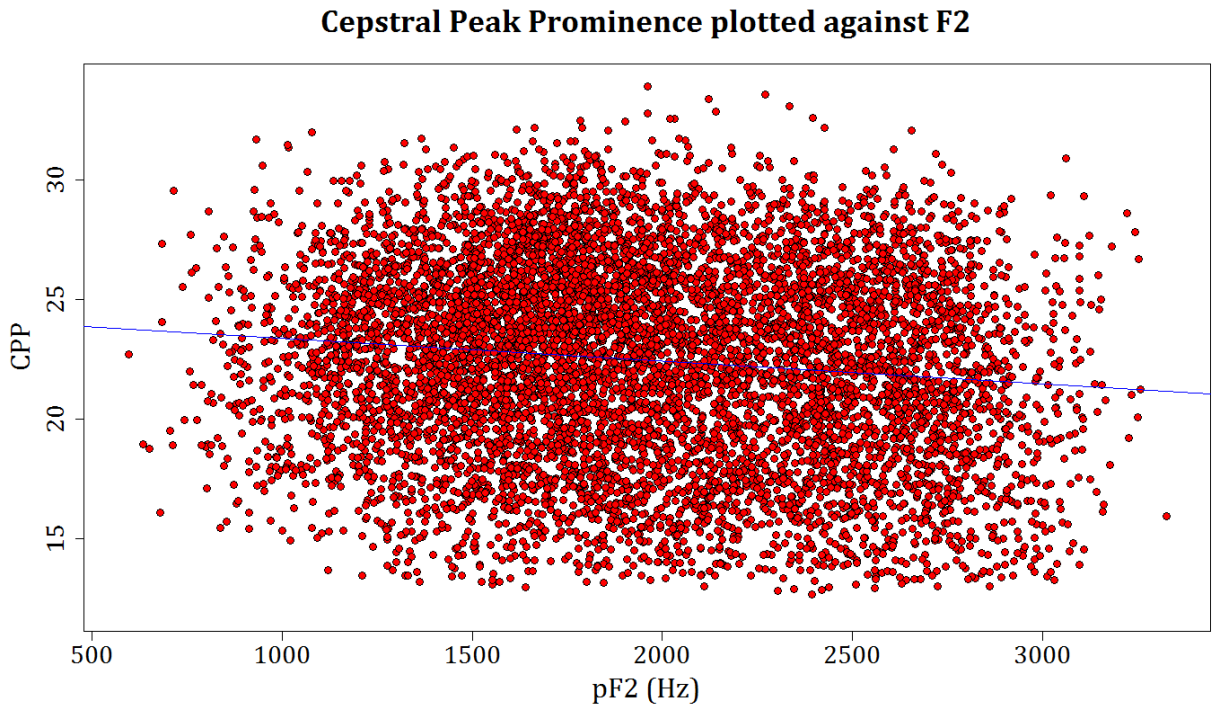

Figure 5.7: Scatterplot of CPP data for the whole sample plotted against pF2 in Hertz (Pearson's $r = -0.12$).

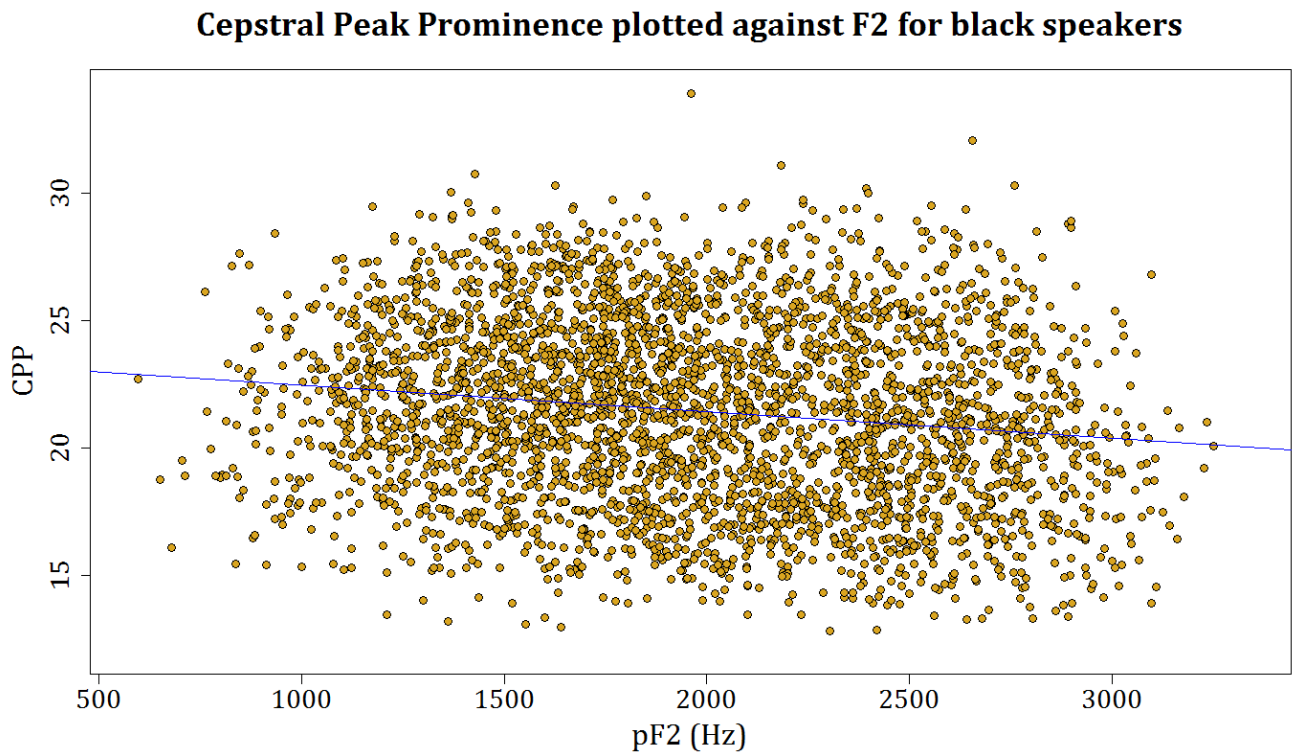

Figure 5.8: Scatterplot of CPP data for black speakers plotted against pF2 in Hertz (Pearson's *r*= −0.153).

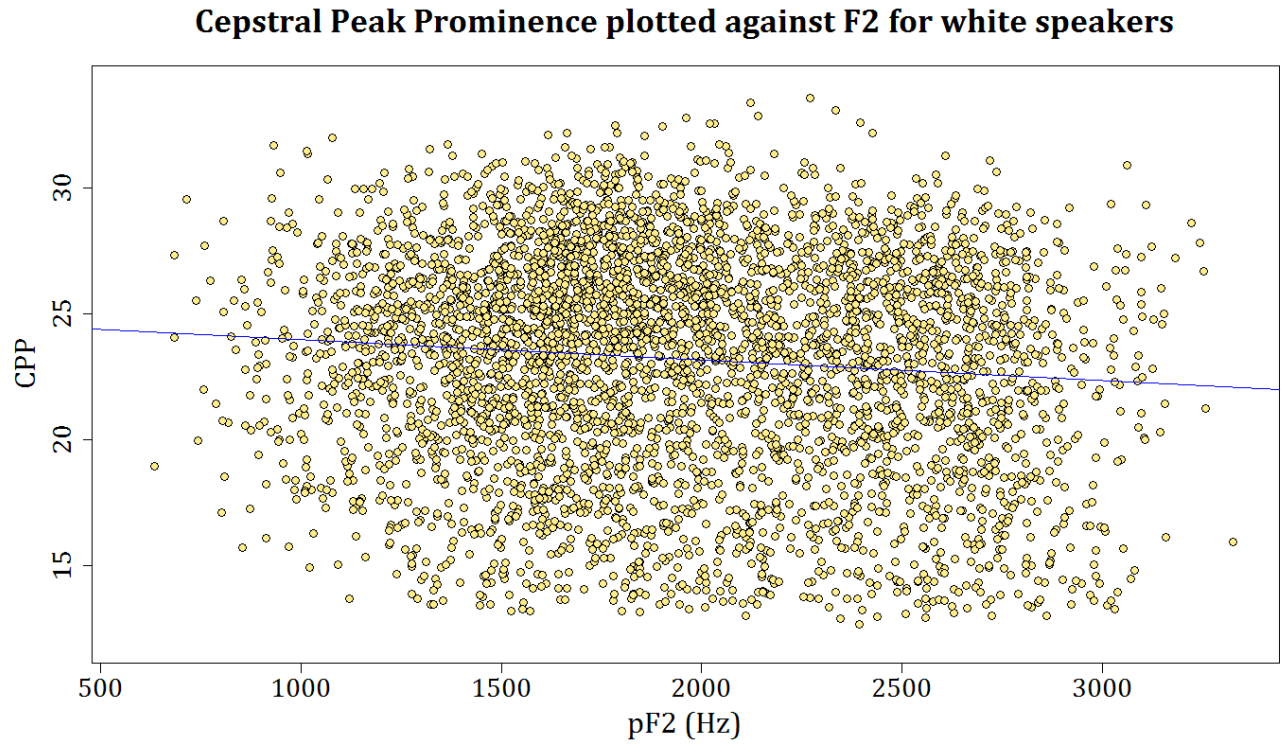

Figure 5.9: Scatterplot of CPP data for white speakers plotted against pF2 in Hertz (Pearson's *r*= −0.097).

As can be seen from figures 5.11, 5.12 and 5.13 below, there is a positive correlation between F1 and CPP. This is expected, given that, as pointed out by Awan, Giovinco and Owens (2012:670.e19), high vowels, that is, vowels with a low F1 are expected to have greater damping of the vocal tract excitation for such vowels which would lead to the damping of the most prominent rahmonic. In addition to this, low vowels (with higher F1 values) are expected to have greater vocal intensity as a result of the openness of the oral cavity in producing some vowels, which would also increase CPP values as a result.

However, it is clear that the correlation is greater for white speakers and the values are still generally higher for white speakers. This may be expected given the higher values for F1 observed for white speakers overall as displayed in figure 5.10 below. However, according to a Wilcoxon rank sum test there is no significant difference between black and white speakers for pF1. If there is a difference between the two groups in terms of F1, even if not significant, this could suggest a difference in terms of articulatory setting rather than a difference in vowel quality, because for this data set all vowel quality differences were averaged out. Possible contributors to this effect could include larynx lowering (as mentioned in chapter 6), as well as certain jaw settings, pharyngealized voice and articulatory settings for pharynx width. Articulatory data would ultimately be needed to evaluate such possibilities.

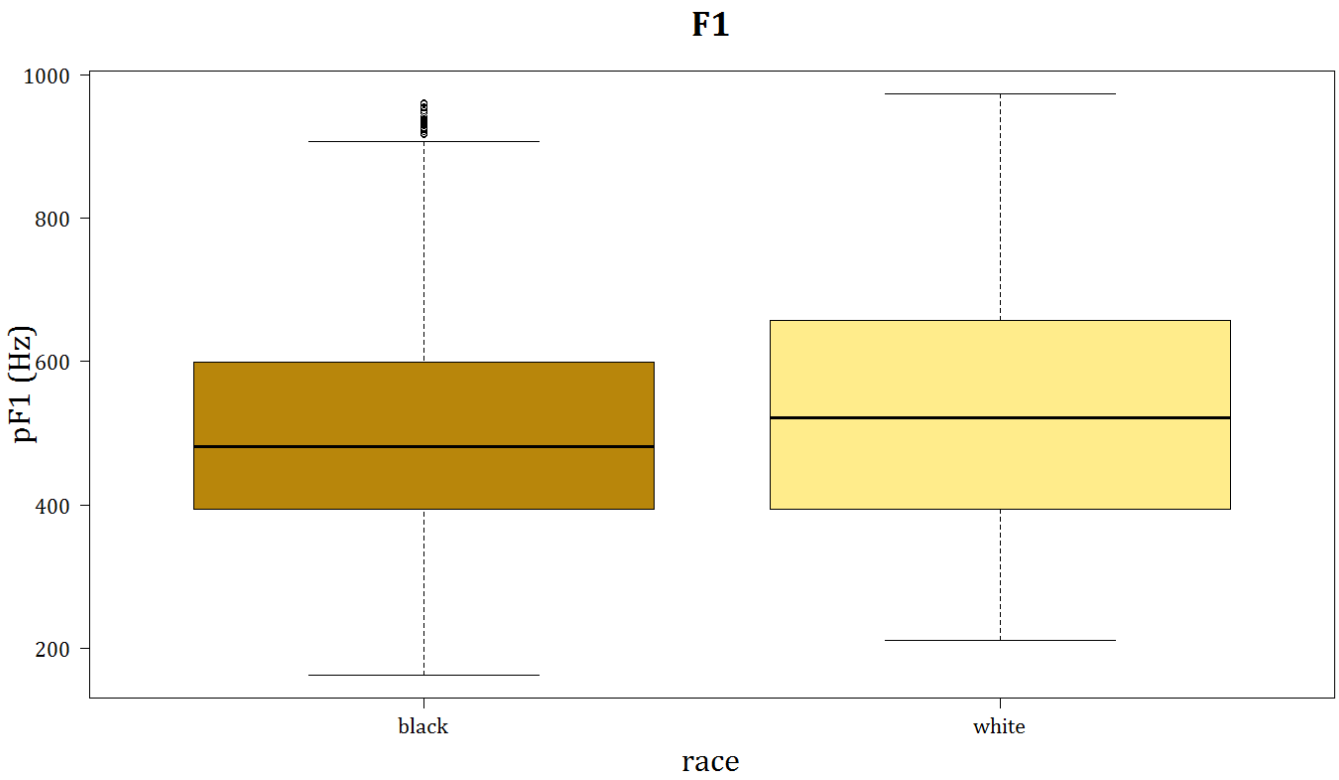

Figure 5.10: Boxplots representing the values of black and white speakers for pF1 measured in Hertz.

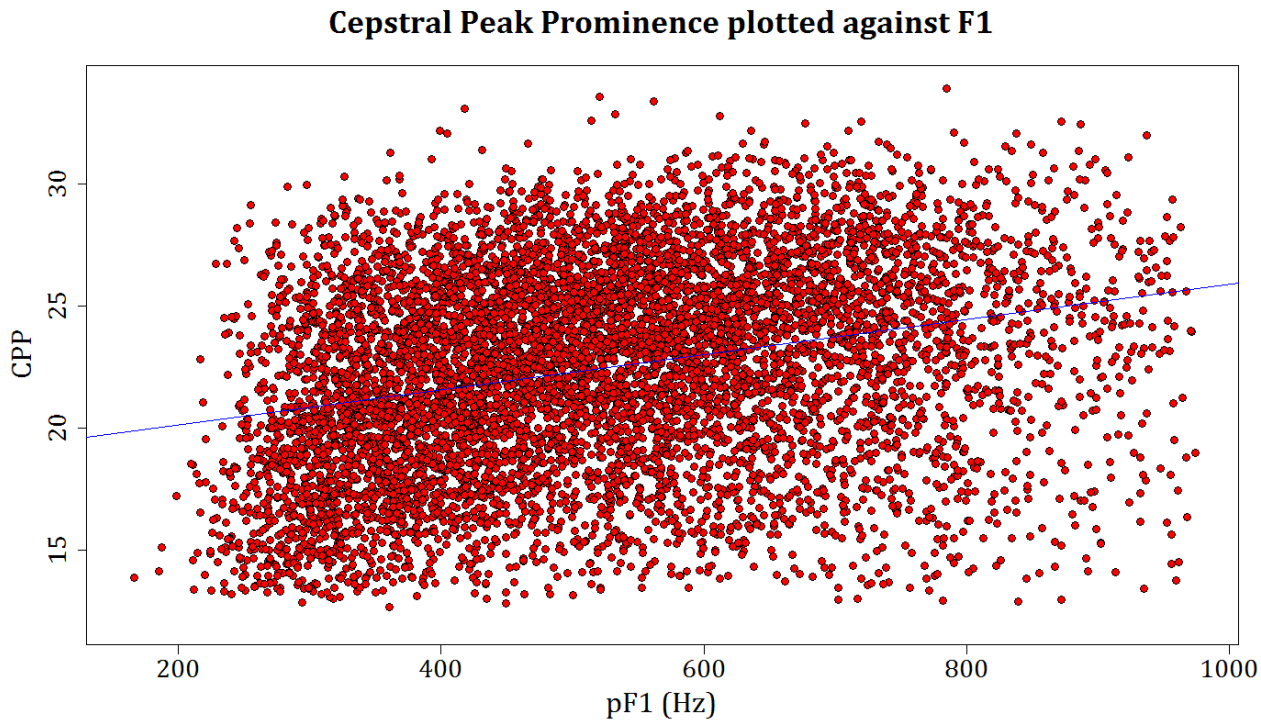

Figure 5.11: Scatterplot of CPP data for the whole sample plotted against pF1 in Hertz (Pearson's *r*= 0.283).

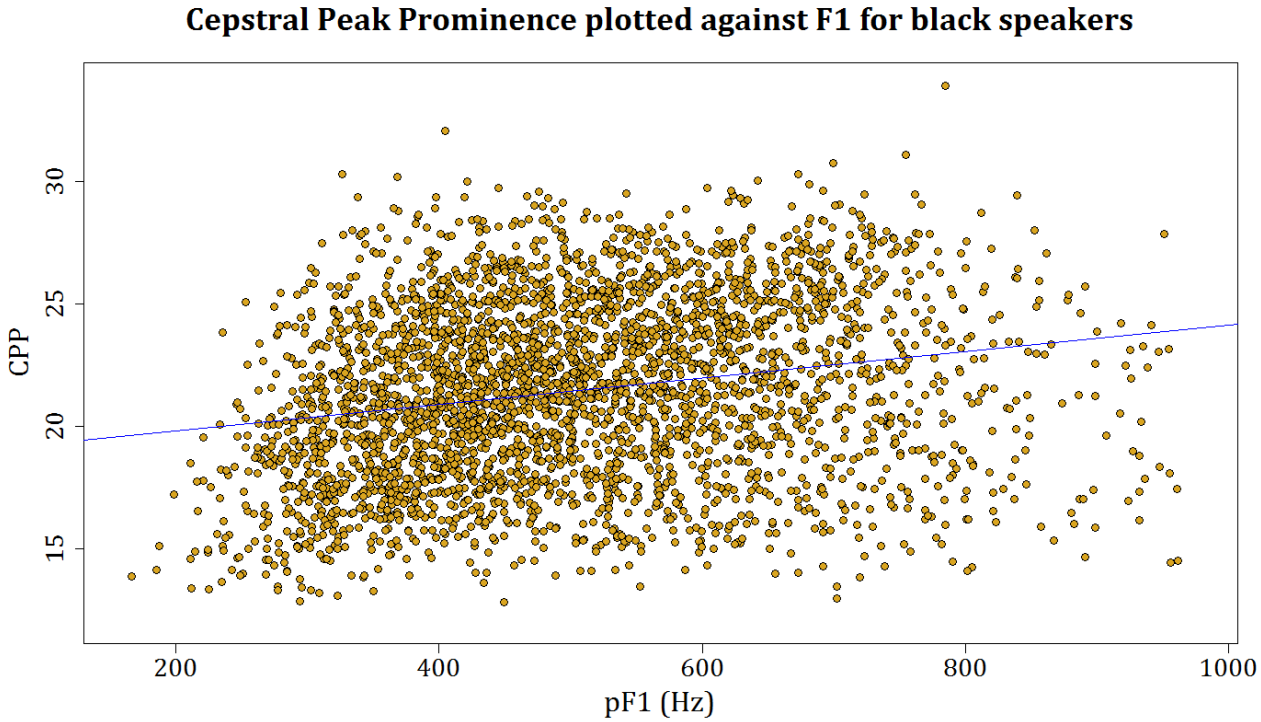

Figure 5.12: Scatterplot of CPP data for black speakers plotted against pF1 in Hertz (Pearson's *r*= 0.219).

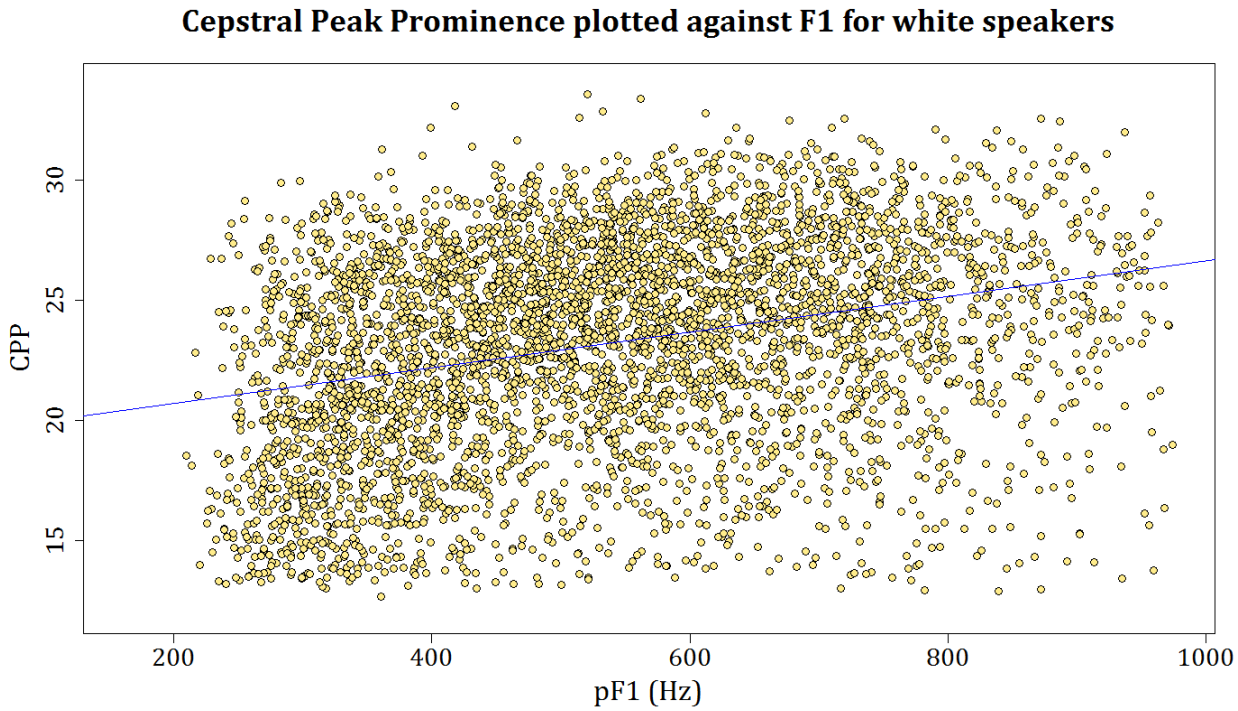

Figure 5.13: Scatterplot of CPP data for white speakers plotted against pF1 in Hertz (Pearson's *r*= 0.296).

As can be observed from figures 5.14, 5.15 and 5.16 below, there is a positive correlation between RMS Energy and CPP, such that the greater the RMS Energy the higher CPP. This would imply that there is a smaller noise component for vowels with increased intensity. This can be easily accounted for, as explained by Awan, Giovinco and Owens (2012:670.e18) who state that medial compression needs to be increased in order to increase loudness of the voice, which in turn results in more effective glottal closure, as well as greater amplitude for vocal fold vibration as well as a generally increased excitation of the vocal tract. There is a tendency for such louder and more intense voices to be produced with a decrease in perturbation, which would thus increase the values for CPP (Awan, Giovinco and Owens 2012:670.e18).

As can be seen from figures 5.15 and 5.16 below, the correlation is greater for black speakers than for white speakers. That is, the decrease in the noise component for increased intensity are greater for black speakers than for white speakers. This may be due, for example, to black speakers using non-modal phonation types to a greater degree for shorter vowels than

white speakers, such that the effect of increasing intensity is more dramatic for black speakers than for white speakers.

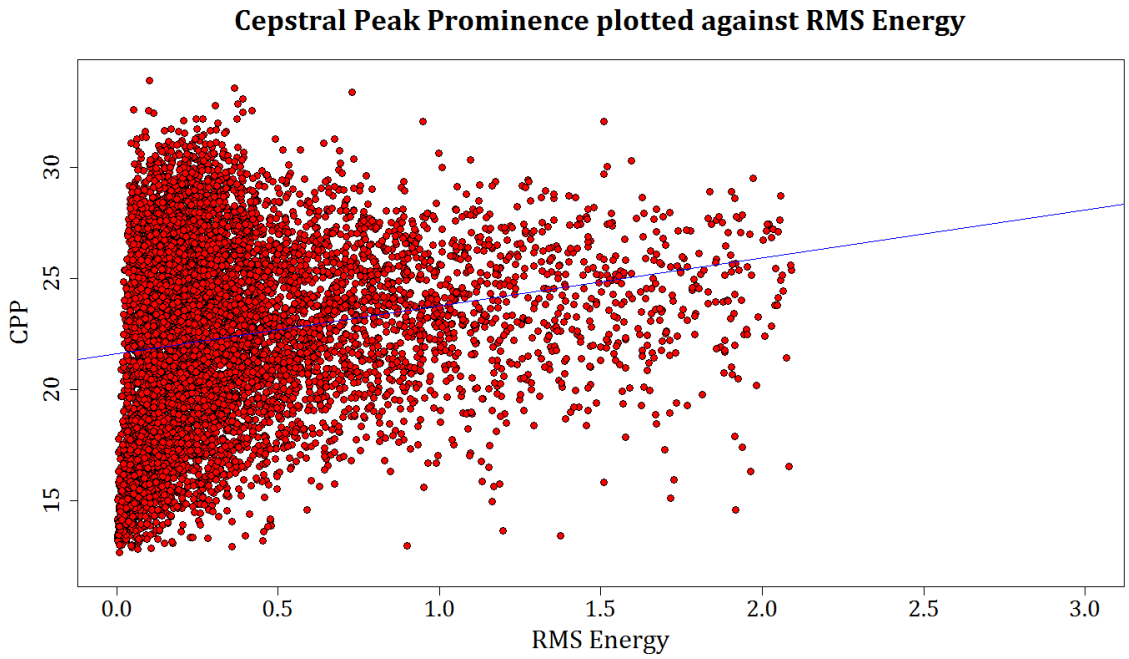

Figure 5.14: Scatterplot of CPP data for the whole sample plotted against RMS energy (Pearson's *r*= 0.202).

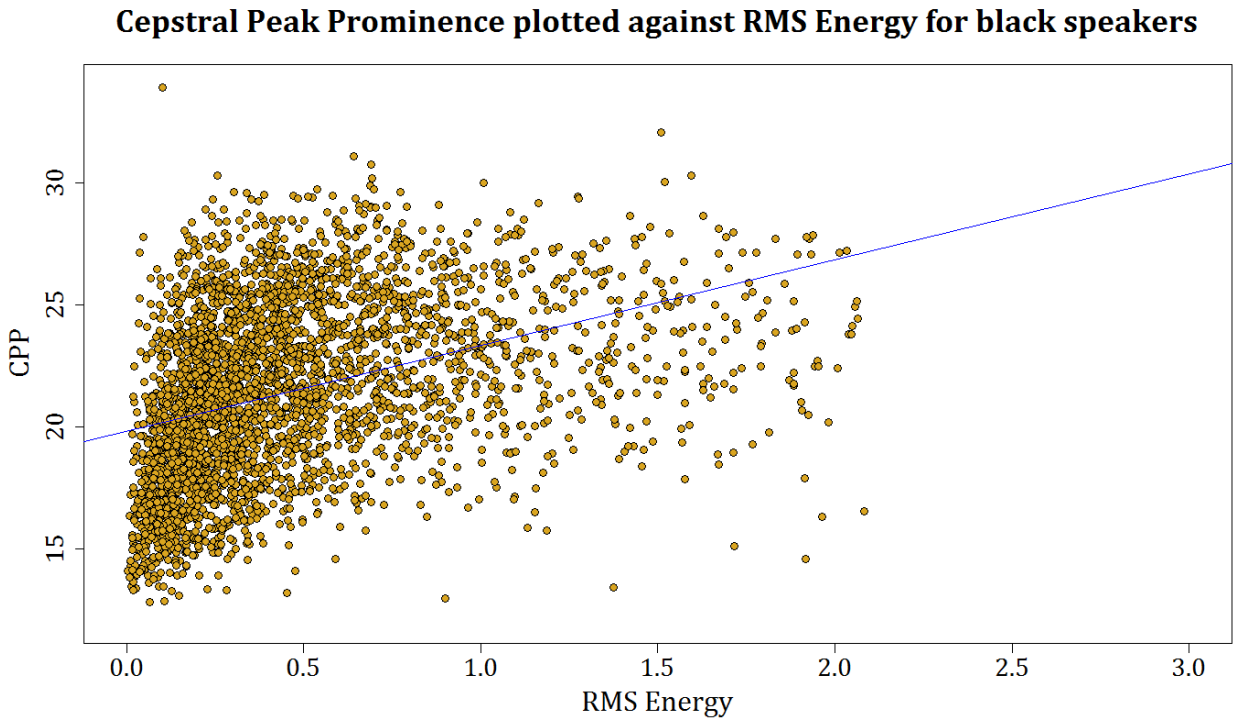

Figure 5.15: Scatterplot of CPP data for black speakers plotted against RMS energy (Pearson's *r*= 0.385).

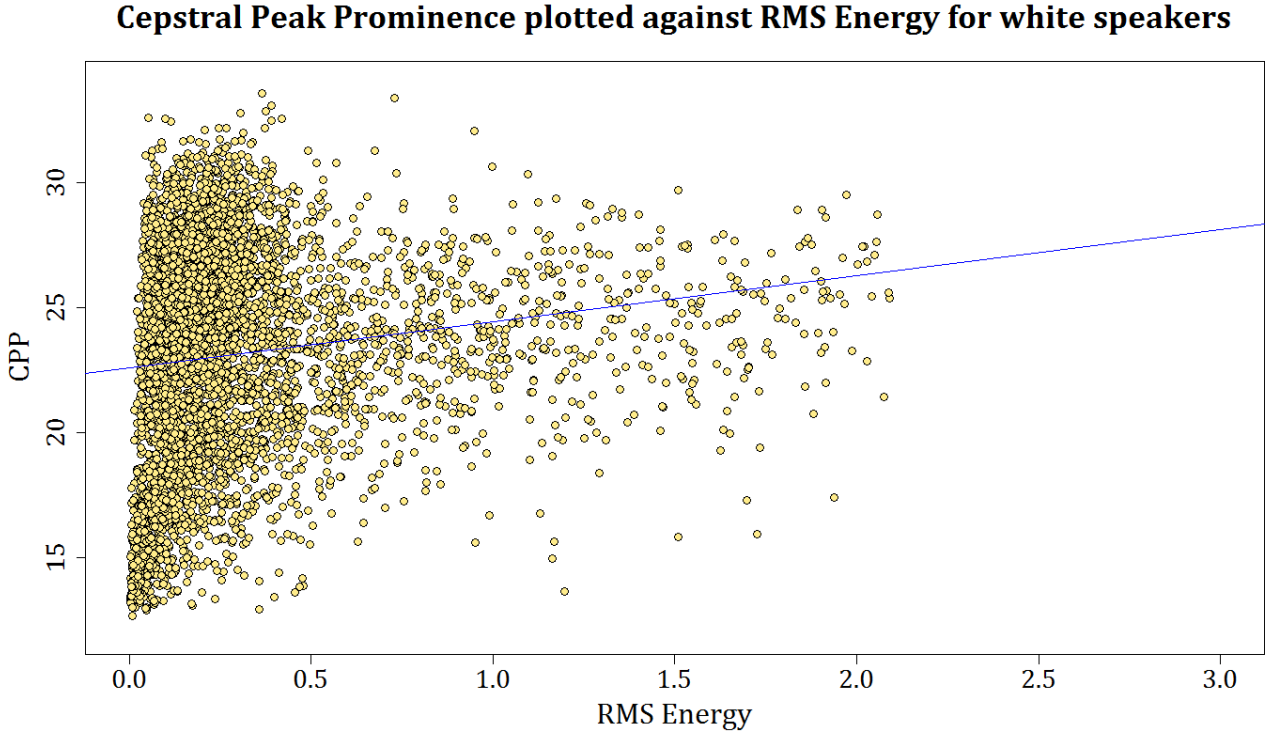

Figure 5.16: Scatterplot of CPP data for white speakers plotted against RMS energy (Pearson's *r*= 0.157).

5.3.3. Summary of Findings for CPP

The Wilcoxon rank sum tests revealed a significant difference between black and white speakers for this measure, with white speakers having higher values overall. This difference is observable from the boxplot comparisons for both the interview data as well as for the sentence data. The results of the linear mixed effects analysis agree with those of the Wilcoxon rank sum tests in that there is a significant effect for ethnicity, with increasing values for white speakers. There are significant effects for all predictors apart from duration. There are also significant interactions between the ethnicity and several other predictors, including fundamental frequency, F1, F2, RMS energy and duration.

Awan, Giovinco and Owens (2012:670.e19) discuss the possibility of vowel-related differences in CPP values. They reason that since F1 and F2 are well separated for high vowels, there will be an general decrease in the amplitude of the signal for such vowels, whereas low vowels by contrast will, due to the proximity of F1 and F2 to one another, have the effect of emphasizing the energy for low vowels overall. Another contributing factor to vowel differences would be that there is a decrease in opening of the oral cavity for high vowels in comparison to that of low vowels, which will increase impedance airflow levels for high vowels, possibly leading to a greater dampening of the vocal tract excitation for such vowels, which would in turn lead to a dampening of the predominant rahmonic and therefore the CPP value for high vowels (Awan, Giovinco and Owens 2012:670.e19). Since the F1 values are greater for white speakers overall in my study, it is possible that at least some of the observed effects could be as the result of this tendency towards higher F1 values for white speakers.

5.4. HARMONICS-TO-NOISE-RATIO

In this section, I discuss the findings for the four harmonics-to-noise ratio measures, namely HNR05, HNR15, HNR25 and HNR35.

5.4.1. HNR05 (the harmonics-to-noise ratio between 0 Hz and 500 Hz)

5.4.1.1. HNR05 Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 5.17 illustrates the boxplots representing the values for the HNR05 data according to auditorily identified phonation type. The highest boxplot, particularly in terms of the interquartile range can be observed for modal voice, followed by those phonation types which involve some level of aspiration noise, such as breathy voice, whisper, whispered creak and whispered vocal fry.

The lowest boxplots, particularly in terms of the interquartile ranges for this measure, are found for the non-composite creak types, namely prototypical creak, vocal fry and harsh/aperiodic voice.

This measure is therefore one of the more consistent differentiators of the auditorily identified phonation types and the pattern of the boxplots for these data are similar to those for the previously discussed measure, namely CPP.

The results are in line with predictions based on prior research as described in the previous chapter with regards to the relation between this voice quality measure and perceptual differences in phonatory quality.

For example, as reported by Keating et al. (2015), lower HNR values would be expected for breathy voice and whisper, due to the presence of noise generated at the glottis for these phonation types and indeed the values of these auditorily identified phonation types do appear lower than that those of modal voice as evidenced by figure 5.17 below.

For auditorily identified voice qualities which are hypothesized to involve a greater degree of irregular fundamental frequency, such as prototypical creak, whispered creak and harsh/aperiodic voice, the HNR values are even lower, as lower HNR values are expected for such phonation types where the periodic excitation is weaker in comparison to glottal noise, as the result of irregular fundamental frequency for these phonation types (Keating et al. 2015).

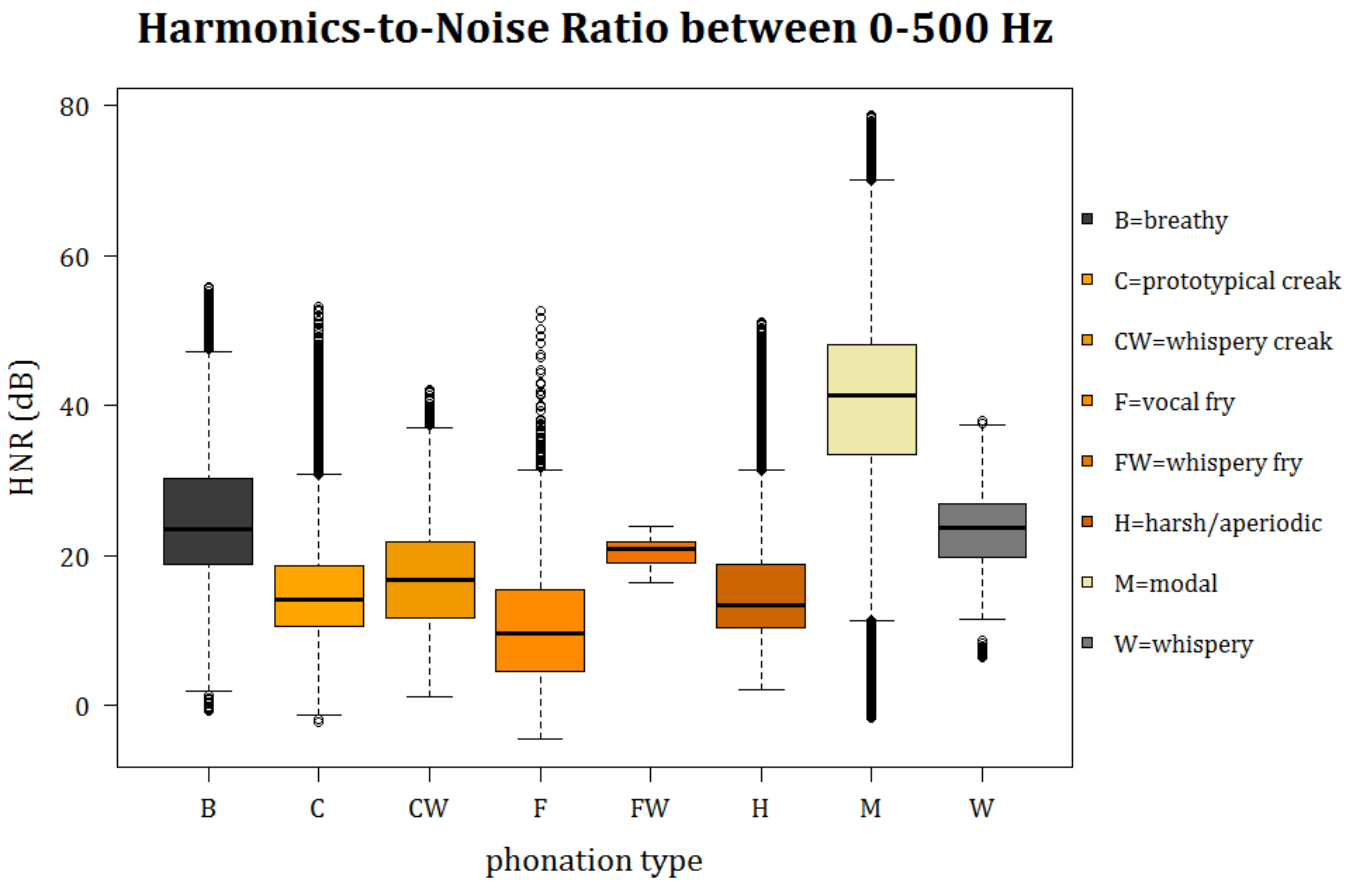

Figure 5.17: Boxplots displaying the values for the HNR05 sentence data for each of the auditorily identified phonation types for all data measurement points.

5.4.1.1. Statistical Analysis for HNR05

According to the results of the Wilcoxon rank sum test, the difference between black and white speakers approaches significance for this measure (W=119, *p*=0.091), for the hypothesis that the values for white speakers are higher than those for black speakers. This difference can also be observed from the boxplot comparison presented in figure 5.18 below.

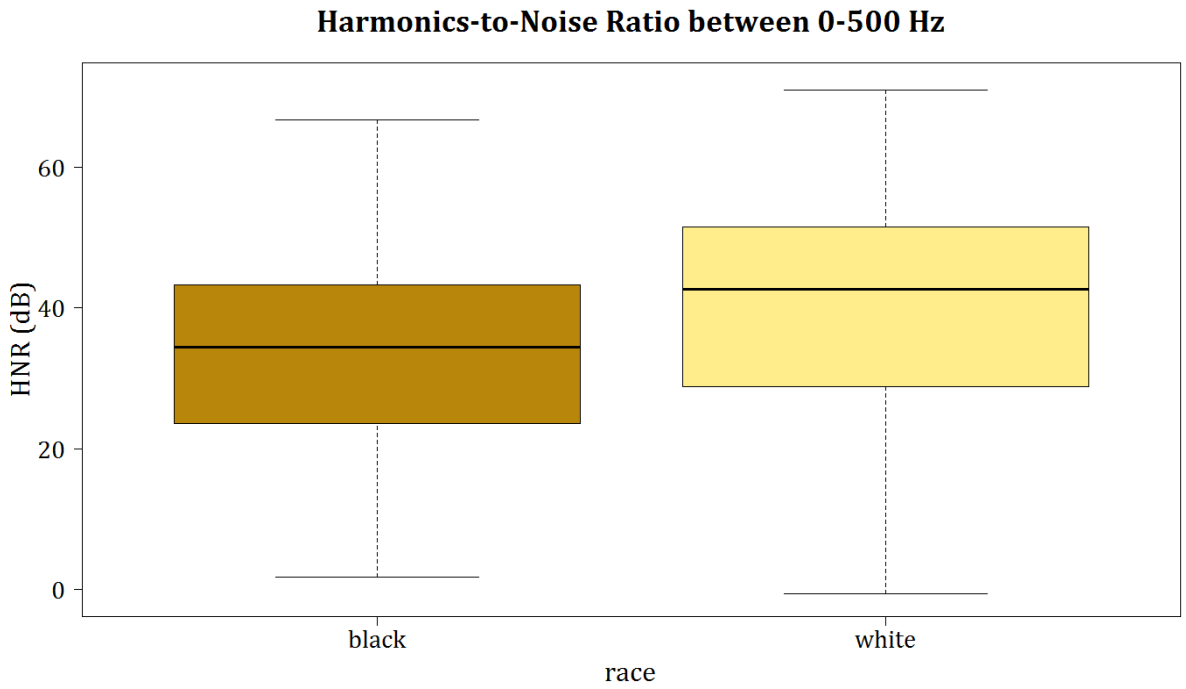

Figure 5.18: Boxplots representing the values of black and white speakers for the HNR05 interview data.

This same pattern is found for the sentence data, as can be seen from figure 5.19 below.

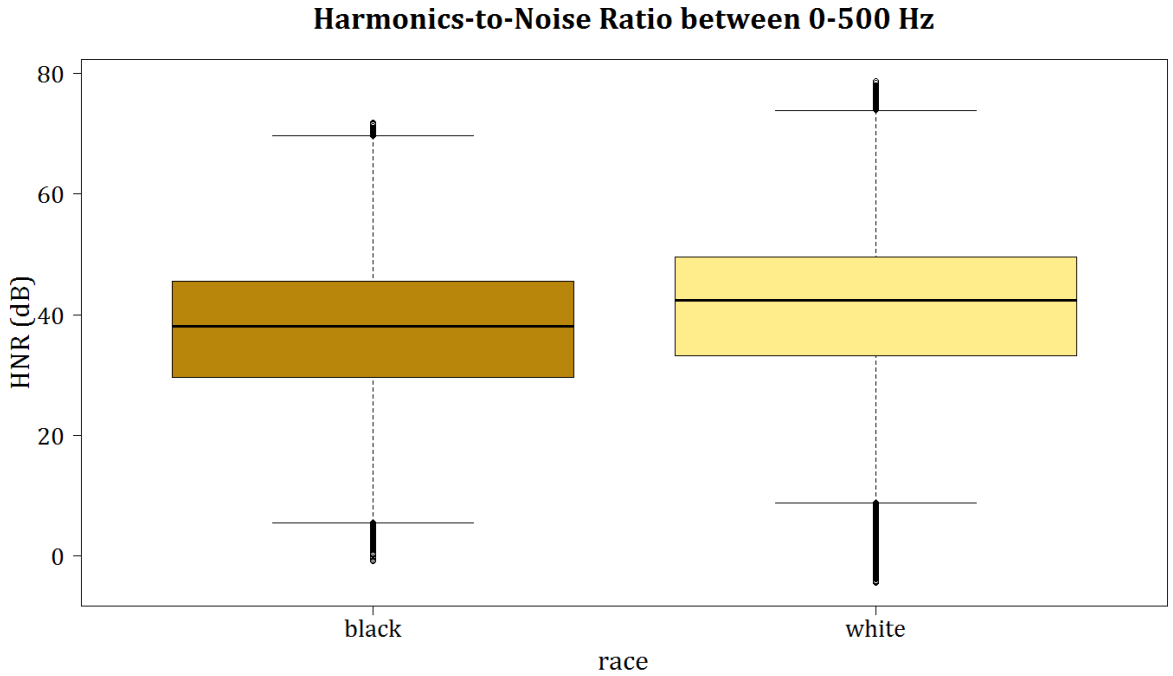

Figure 5.19: Boxplots representing the values for the HNR05 sentence data for all data measurement points according to ethnicity.

5.4.1.1.1. Interview Data

There is however no significant effect for ethnicity according to the results of the linear mixed effects analysis ($X^2(1)= 1.0644$;$p=0.302$), with an increase of 2.229 dB ±2.1368 (standard errors) in HNR05 for white speakers.

However the results of all of the other predictors are highly significant for the interview data. There is a significant effect for logpF0 ($X^2(1)= 3996.4$;$p<0.001$), a 1% increase in pF0 increasing HNR05 by 0.221 dB ±0.3070 (standard errors) as well as for logEnergy ($X^2(1)= 190.88$;$p<0.001$), a 1% increase in RMS Energy increasing HNR05 by 0.019 dB ±0.1376 (standard errors), logpF1 ($X^2(1)= 24.463$;$p<0.001$), a 1% increase in pF1 increasing HNR05 by 0.02 dB ±0.4106 (standard errors), logpF2 ($X^2(1)= 92.892$;$p<0.001$), a 1% increase in pF2 decreasing HNR05 by 0.044 dB ±0.4499 (standard errors), logduration ($X^2(1)= 80.176$;$p<0.001$), a 1% increase in duration increasing HNR05 by 0.024 dB ±0.2650 (standard errors) as well as for speaker ($X^2(3)= 2505.5$;$p<0.001$).

There are however also some interactions between ethnicity and other predictors for this measure. The interaction between ethnicity and logEnergy is significant ($X^2(1)= 10.453$;$p=0.001$) and the interaction between this variable and logpF0 approaches significance ($X^2(1)= 2.8459$;$p=0.091$). The interaction between ethnicity and duration ($X^2(1)= 1.1273$;$p<0.001$) is also significant.

5.4.1.1.2. Sentence Data Linear Mixed Effects Analysis Results

As for the interview data, there is likewise no significant effect for the sentence data ($X^2(1)=1.1719$, $p=0.279$), although the effect for the interaction between ethnicity and vowel is highly significant ($X^2(2)=2022.5$, $p<0.001$).

5.4.1.2. Summary of the Findings for HNR05

There are differences approaching significance between black and white speakers for this measure according to the results of the Wilcoxon rank sum tests, with white speakers having higher values than black speakers. This difference can be seen from the boxplot comparisons for both the interview data as well as for the sentence data. There is no significant difference for ethnicity according to the results of the linear mixed effects analysis, with an increase in HNR05 values for white speakers. There are significant effects for all of the other predictors. There is

also a significant interaction between ethnicity and RMS energy as well as between ethnicity and duration, while the interaction between ethnicity and fundamental frequency approaches significance.

5.4.2. HNR15 (the harmonics-to-noise ratio between 0 Hz and 1500 Hz)

5.4.2.1. HNR15 Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 5.20 displays the boxplots representing the values for the HNR15 sentence data grouped according to auditorily identified phonation type. In general, this measure displays a similar pattern to that of HNR05 described above, in that the phonation type with the highest boxplot for this measure particularly in terms of the interquartile range is that of modal voice, followed by those phonation types which are expected to involve some degree of aspiration, such as breathy voice, whispery creak and whisper. Values for whispery vocal fry are found at the lower end of this range. The lowest boxplots, particularly in terms of interquartile ranges are found for the creak types prototypical creak, vocal fry and whispery vocal fry.

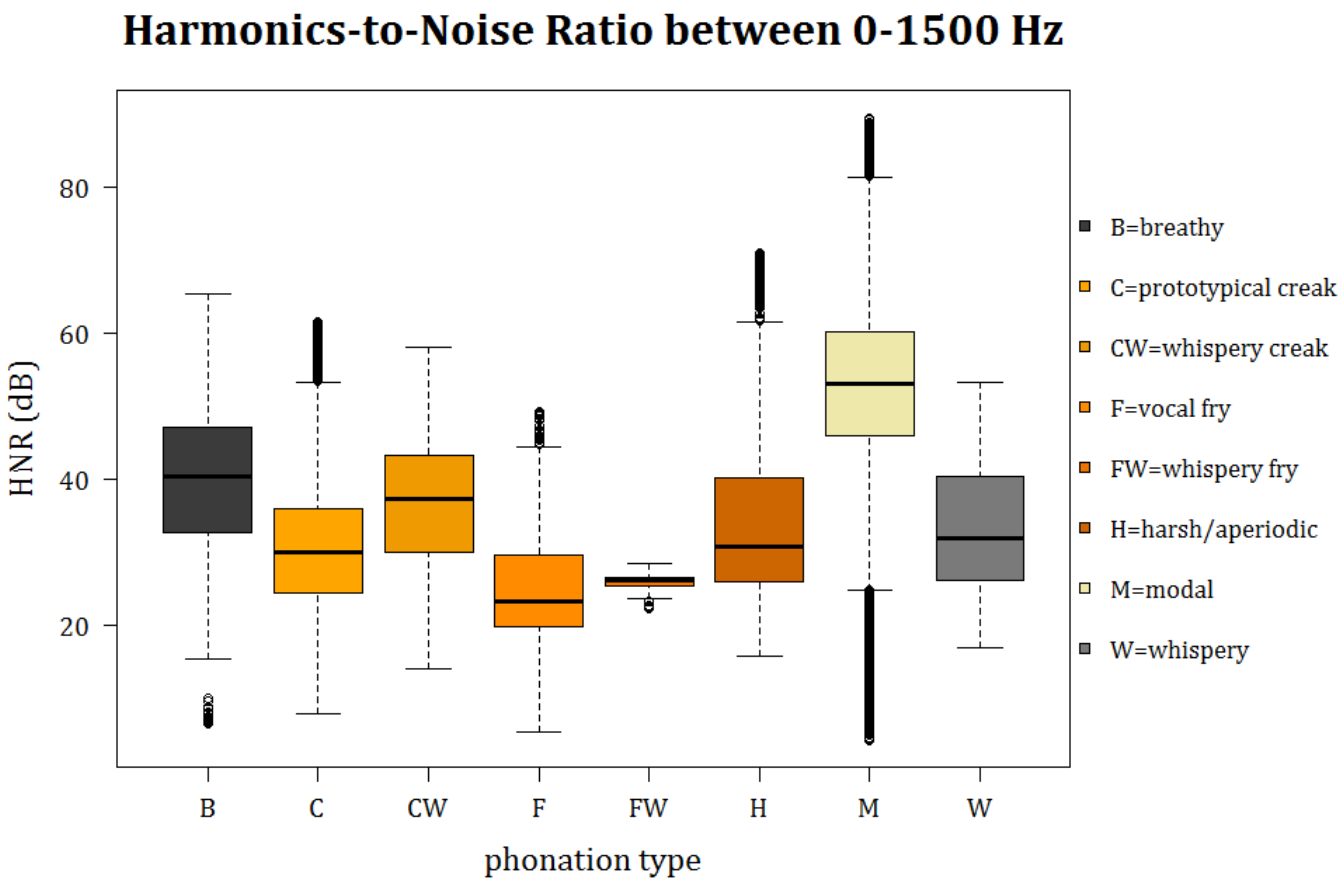

Figure 5.20: Boxplots displaying the values for the HNR15 sentence data for each of the auditorily identified phonation types for all data measurement points.

5.4.2.2. Statistical Analysis for HNR15

The Wilcoxon rank sum test results reveal no significant effect for ethnicity. However white speakers exhibit higher values as can be observed from the boxplot comparison presented in figure 5.21 below.

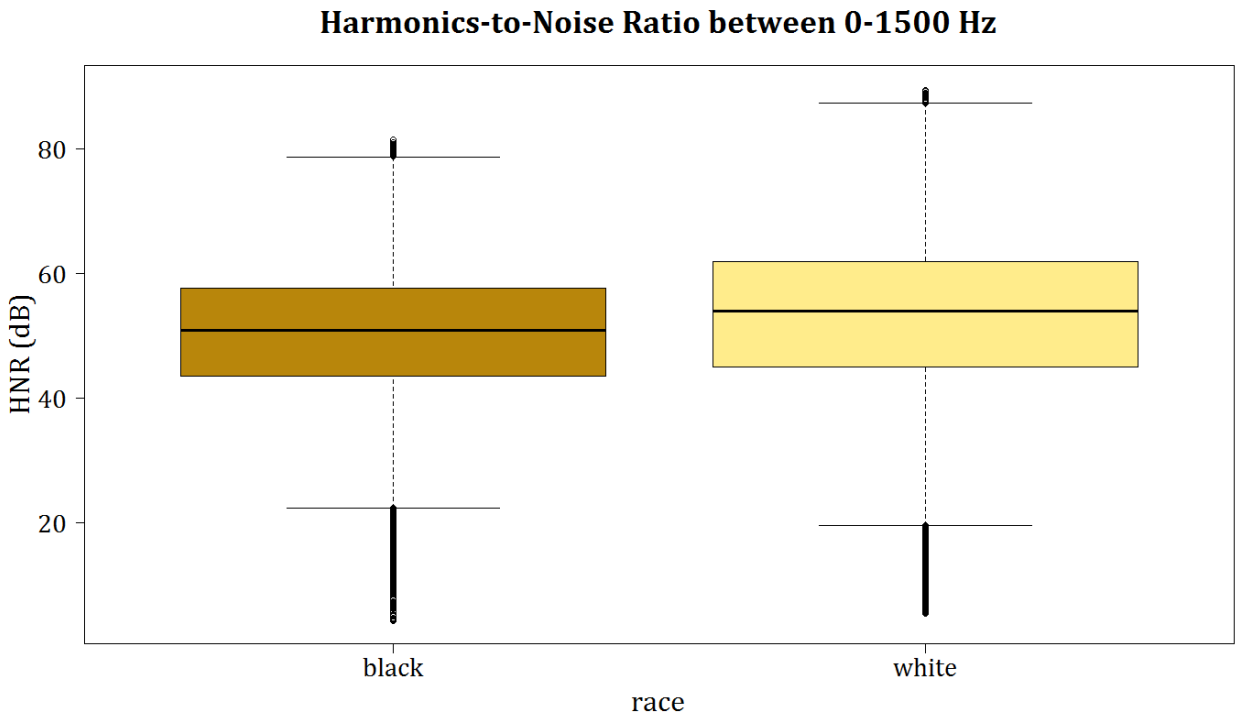

Figure 5.21: Boxplots representing the values of black and white speakers for the HNR15 interview data.

The same pattern can be seen from the sentence data as displayed in figure 5.22 below.

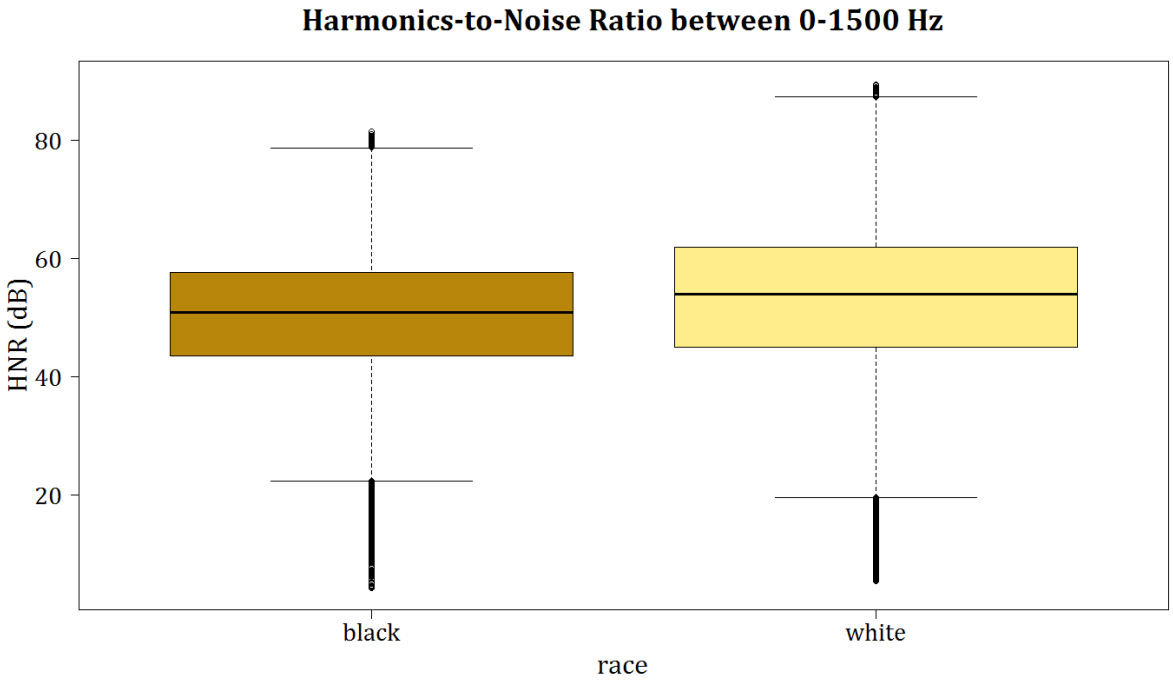

Figure 5.22: Boxplots representing the values for the HNR15 sentence data for all data measurement points according to ethnicity.

5.4.2.2.1. Interview Data

There is however no significant effect for ethnicity for this measure according to the results of the linear mixed effects analysis ($X^2(1)= 1.0344$;$p=0.309$) for the interview data, with an increase of 1.96 dB ±1.9073 (standard errors) in HNR15 for white speakers.

However, for all of the other predictors, for the interview data there are highly significant effects for this measure. The effect for logpF0 is highly significant ($X^2(1)= 3569.1$;$p<0.001$), a 1% increase in pF0 increasing HNR15 by 0.192 dB ±0.2852 (standard errors), as is the effect for logEnergy ($X^2(1)= 83.882$;$p<0.001$), a 1% increase in RMS Energy increasing HNR15 by 0.012 dB ±0.1277 (standard errors), logpF1 ($X^2(1)= 578.35$;$p<0.001$), a 1% increase in pF1 decreasing HNR15 by 0.095 dB ±0.3803 (standard errors), logduration ($X^2(1)= 59.9$;$p<0.001$), a 1% increase in duration increasing HNR15 by 0.019 dB ±0.2461 (standard errors), logpF2 ($X^2(1)= 84.739$;$p<0.001$), a 1% increase in pF2 increasing HNR15 by 0.039 dB ±0.4166 (standard errors) and speaker ($X^2(3)= 2372.9$;$p<0.001$).

There is only one significant interaction with ethnicity for this variable, namely for logEnergy ($X^2(1)= 4.9919; p=0.025$), while the interaction between ethnicity and logpF0 approaches significance ($X^2(1)= 3.1541; p=0.076$).

5.4.2.2.2. Sentence Data Linear Mixed Effects Analysis Results

As for the interview data, for the sentence data, there is no significant effect for ethnicity ($X^2(1)=1.3264$, $p=0.249$), although there is a significant effect for the interaction between ethnicity and vowel category ($X^2(2)=1402.9$, $p<0.001$).

5.4.2.3. Summary of the Findings for HNR15

There is no significant difference between black and white speakers for this measure according to the results of the Wilcoxon rank sum test. White speakers nevertheless exhibit higher values than black speakers which is evident from the boxplots for both the interview data as well as for the sentence data. There is no significant effect for ethnicity according to the results of the linear mixed effects analysis, with an increase in HNR15 values for white speakers. There are significant effects for all of the other predictors. There is also a significant interaction between ethnicity and RMS energy and an interaction which approaches significance between ethnicity and fundamental frequency.

5.4.3. HNR25 (the harmonics-to-noise ratio between 0 Hz and 2500 Hz)

5.4.3.1. HNR25 Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 5.23 illustrates the boxplots representing the values for the HNR25 sentence data for each of the auditorily identified phonation types for all of the data measurement points. The pattern of the boxplots for HNR25 is similar to that of the other HNR measures reported above, with the exception that whisper appears to have a lower boxplot than for the other HNR measures reported thus far and whispery vocal fry also has a particularly low interquartile range. Modal voice exhibits the highest interquartile range, being followed by those auditorily identified voice qualities such as breathy voice and whispery creak which are expected to involve some degree of aspiration. With the exception of whisper, the lowest interquartile ranges are found for phonation types involving creak.

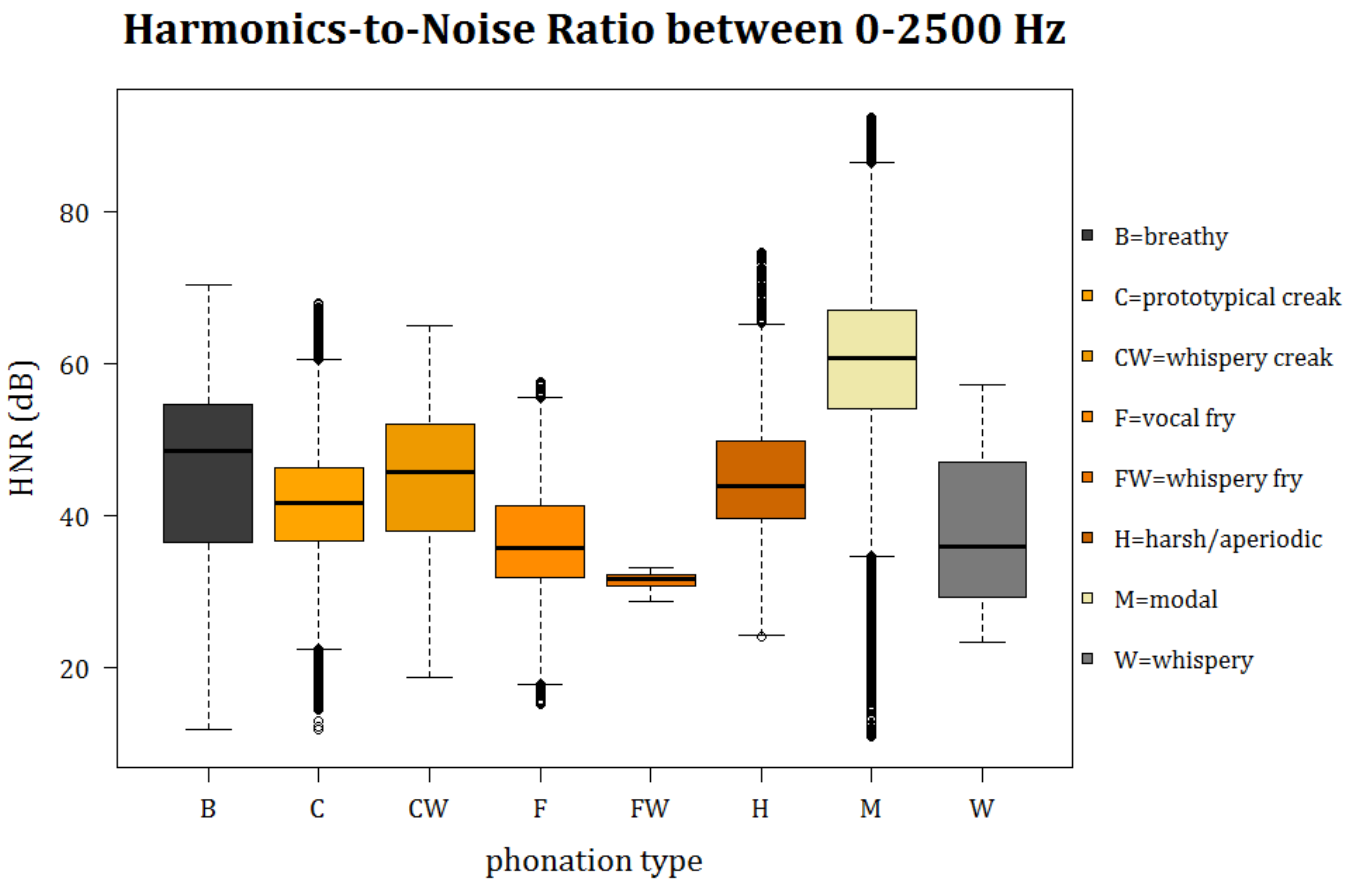

Figure 5.23: Boxplots displaying the values for the HNR25 sentence data for each of the auditorily identified phonation types for all data measurement points.

5.4.3.2. Statistical Analysis for HNR25

The Wilcoxon rank sum test results reveal no significant difference between white and black speakers for this variable. White speakers nevertheless exhibit higher values than black speakers as can be observed from the boxplot comparison of the values displayed in figure 5.24 below.

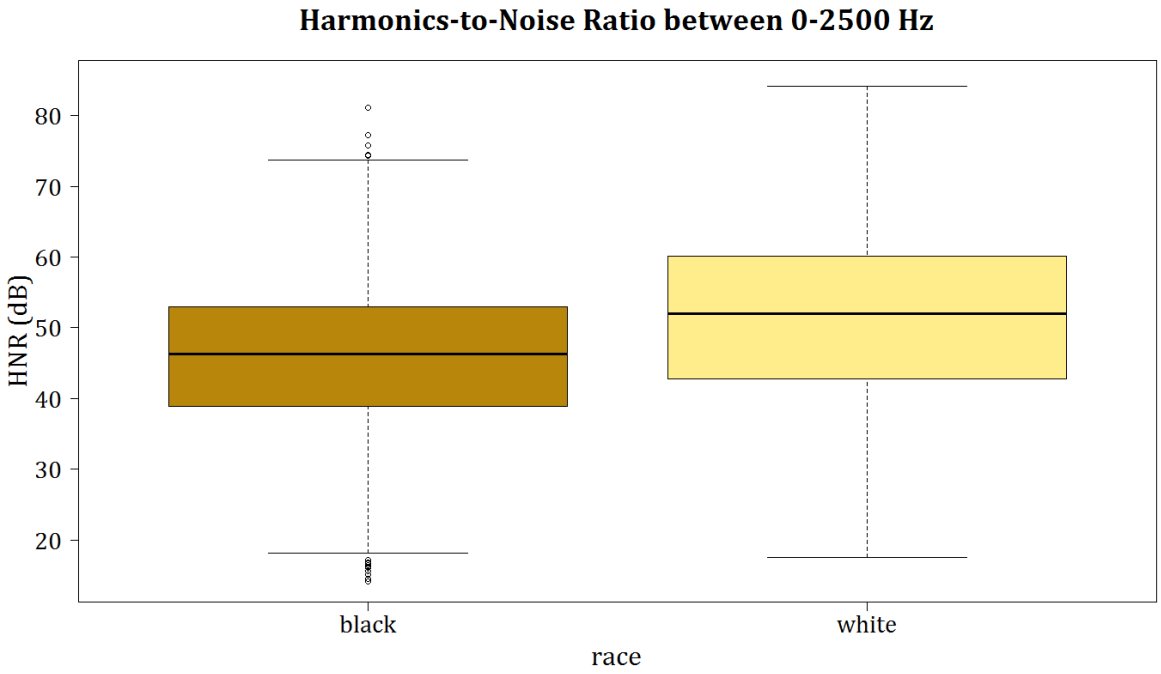

Figure 5.24: Boxplots representing the values of white and black speakers for the HNR25 interview data.

A similar pattern is found for the sentence data, displayed in figure 5.25 below.

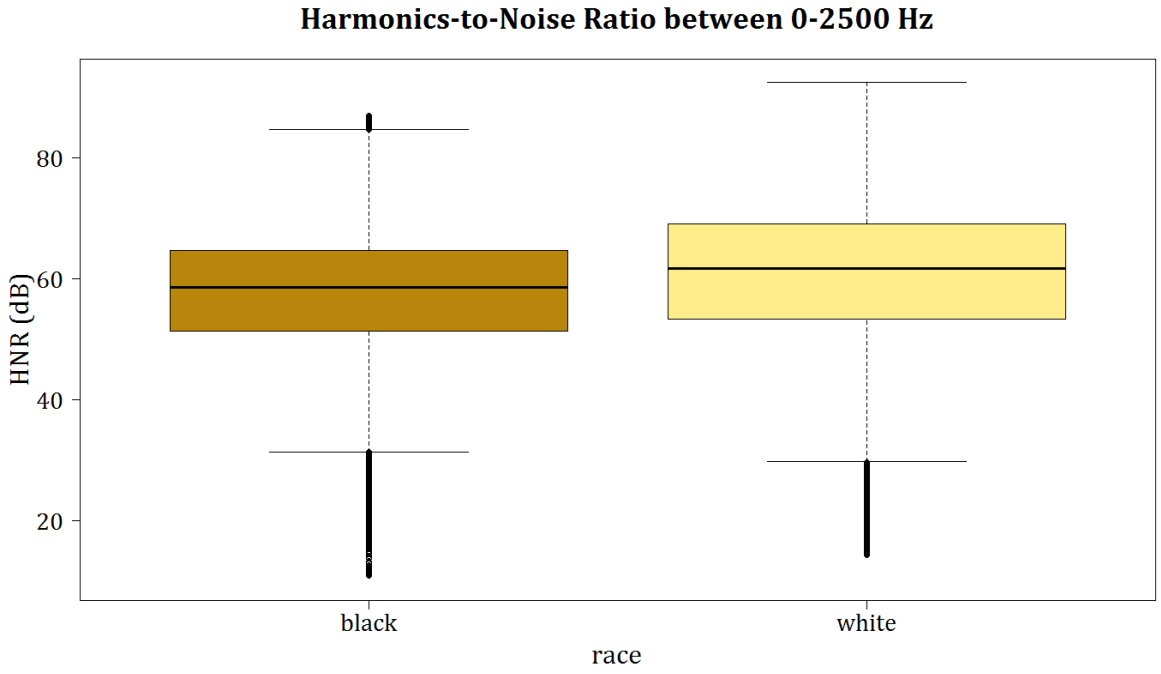

Figure 5.25: Boxplots representing the values for the HNR25 sentence data for all data measurement points according to ethnicity.

5.4.3.2.1. Interview Data

There is however no significant effect for ethnicity based on the results of the linear mixed effects model regression analysis ($X^2(1)= 1.3636$;$p=0.243$), with an increase of 1.92 dB ±1.6230 (standard errors) in HNR25 for white speakers.

There are however significant effects for the other predictors for the interview data. There is a significant effect for logpF0 ($X^2(1)= 3364.3$;$p<0.001$), a 1% increase in pF0 increasing HNR25 by 0.167 dB ±0.2565 (standard errors), as well as for logEnergy ($X^2(1)= 8.1439$;$p=0.004$), a 1% increase in RMS Energy increasing HNR25 by 0.003 dB ±0.1151 (standard errors), logpF1 ($X^2(1)= 747.98$;$p<0.001$), a 1% increase in pF1 decreasing HNR25 by 0.099 dB ±0.3468 (standard errors), logpF2 ($X^2(1)= 6.6179$;$p=0.01$), a 1% increase in pF2 increasing HNR25 by 0.01 dB ±0.3806 (standard errors), logduration ($X^2(1)= 102.17$;$p<0.001$), a 1% increase in duration increasing HNR25 by 0.023 dB ±0.2220 (standard errors) and speaker ($X^2(3)= 2136.5$;$p<0.001$).

There is a significant interaction between ethnicity and logpF0 ($X^2(1)= 8.5667$;$p=0.003$), as well as for logEnergy ($X^2(1)= 7.4036$;$p=0.007$) and logpF1 ($X^2(1)= 9.7968$;$p=0.002$).

5.4.3.2.2. Sentence Data Linear Mixed Effects Analysis Results

As for the interview data, there is no significant effect for ethnicity for the sentence data ($X^2(1)=1.8875$, $p=0.17$), although there is a highly significant effect for the interaction between ethnicity and vowel ($X^2(2)=1796.7$, $p<0.001$).

5.4.3.3. Summary of Findings for HNR25

There is a significant difference between black and white speakers for this measure according to the Wilcoxon rank sum test results, such that white speakers have higher values than black speakers. This is evident from the boxplot comparisons for both interview and sentence data. There is however no significant effect for ethnicity based on the results of the linear mixed effects analysis. There are significant effects for all other predictors. There is a significant interaction between ethnicity and fundamental frequency, ethnicity and RMS energy and between ethnicity and first formant frequency.

5.4.4. HNR35 (the harmonics-to-noise ratio between 0 Hz and 3500 Hz)

5.4.4.1. HNR35 Sentence Data and the Auditorily Identified Phonation Types

The following figure, figure 5.26 displays the boxplots representing the values for the HNR35 (harmonics-to-noise ratio in the 0-3500Hz range) sentence data according to auditorily identified phonation types. This patterning of the boxplots is fairly similar to that of the other HNR measures discussed above, particularly to the values for HNR25, although for HNR35, the distinction between the different phonation types as auditorily identified is less clear. However, modal voice is still the phonation type exhibiting the highest interquartile range, while whisper displays the lowest.

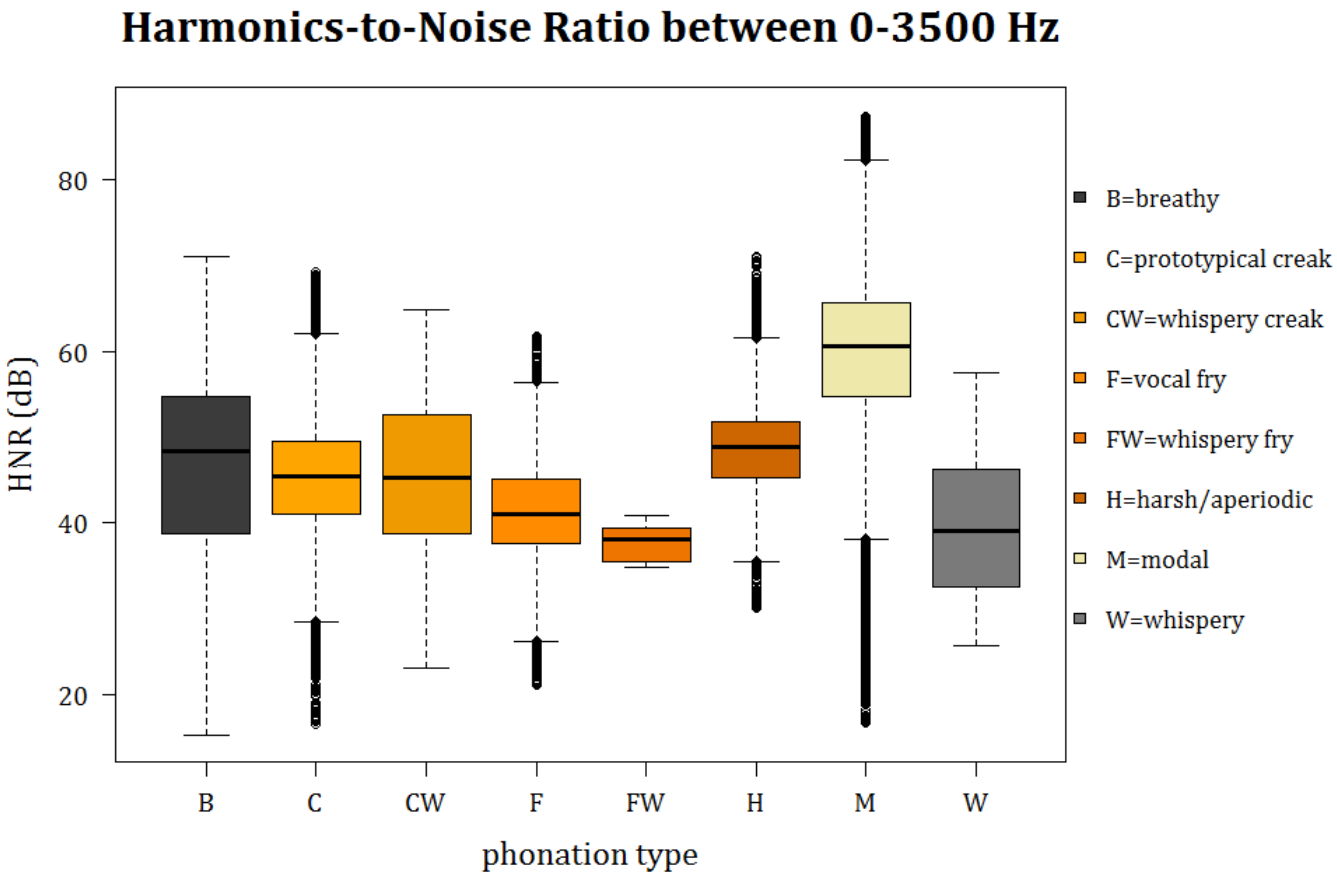

Figure 5.26: Boxplots displaying the values for the HNR35 sentence data for each of the auditorily identified phonation types for all data measurement points.

5.4.4.2. Statistical Analysis for HNR35

The results of the Wilcoxon rank sum test indicate a difference approaching significance between black and white speakers (W=117, *p*=0.081), for the alternative hypothesis that the values for white speakers are higher than those for black speakers. This difference can also be observed from the boxplot comparison provided in figure 5.27 below, where white speakers show higher values for this measure than black speakers.

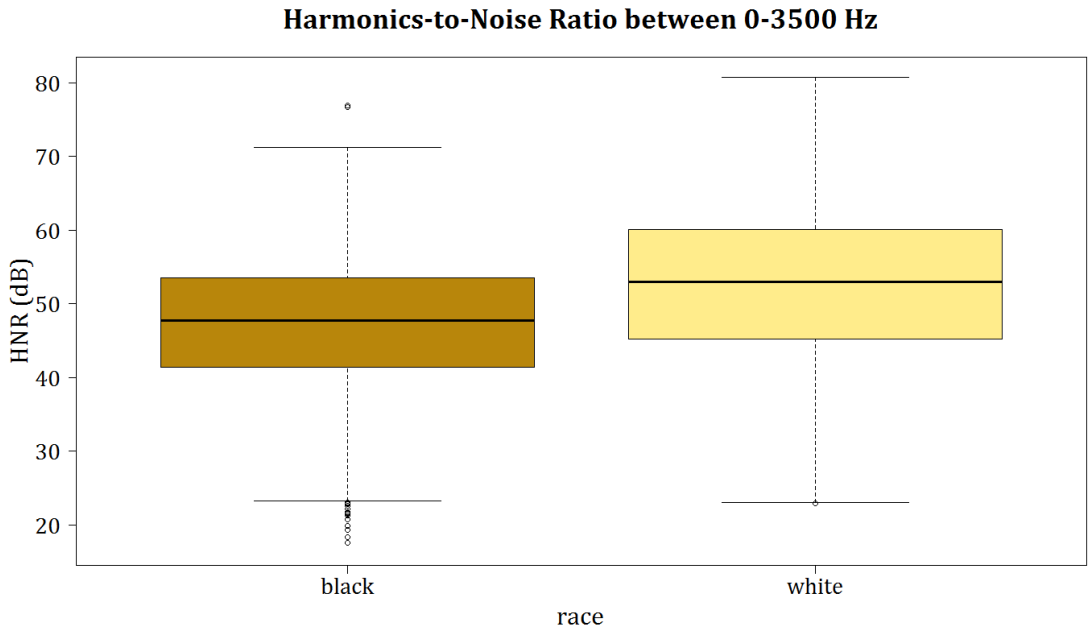

Figure 5.27: Boxplots representing the values for black and white speakers for the HNR35 interview data.

The same pattern can also be seen for the sentence data as displayed in figure 5.28 below.

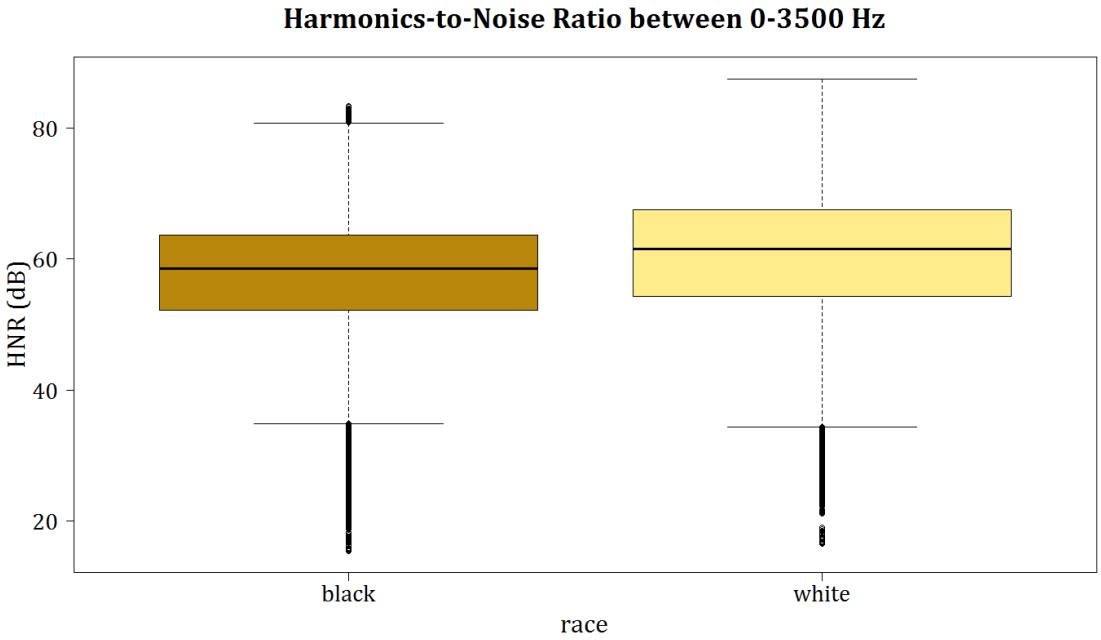

Figure 5.28: Boxplots representing the values for the HNR35 sentence data for all data measurement points according to ethnicity.

5.4.4.2.1. Interview Data

There is no significant effect for ethnicity alone as a variable for this measure, based on the results of the linear mixed effects analysis ($X^2(1)= 2.5486$;$p=0.11$), with an increase of 2.203 dB ±1.3495 (standard errors) for white speakers.

All other predictors exhibit significant effects for this measure. These include logpF0 ($X^2(1)= 3153.9$;$p<0.001$), a 1% increase in pF0 increasing HNR35 by 0.143 dB ±0.2281 (standard errors), logEnergy ($X^2(1)= 20.147$;$p<0.001$), a 1% increase in RMS Energy decreasing HNR35 by 0.005 dB ±0.1024 (standard errors), logpF1 ($X^2(1)= 644.51$;$p<0.001$), a 1% increase in pF1 decreasing HNR35 by 0.082 dB ±0.3097 (standard errors), logpF2 ($X^2(1)= 31.634$;$p<0.001$), a 1% increase in pF2 decreasing HNR35 by 0.019 dB ±0.3402 (standard errors), logduration ($X^2(1)= 91.237$;$p<0.001$), a 1% increase in duration increasing HNR35 by 0.019 dB ±0.1976 (standard errors) as well as speaker ($X^2(3)= 1848.5$;$p<0.001$).

There is a highly significant interaction between ethnicity and logpF1 ($X^2(1)= 11.433$;$p=0.001$), as well as between ethnicity and logpF0 ($X^2(1)= 16.53$;$p<0.001$). There is also a significant interaction between ethnicity and logEnergy ($X^2(1)= 9.3139$;$p=0.002$).

5.4.4.2.2. Sentence Data Linear Mixed Effects Analysis Results

There is no significant effect for ethnicity for the sentence data (X^2 (1)=0, p=1), although the interaction between ethnicity and vowel for the sentence data is highly significant (X^2 (2)=3074.2, p<0.001).

5.4.4.3. Summary of Findings for HNR35

The results of the Wilcoxon rank sum tests reveal no significant difference between black and white speakers. White speakers nevertheless exhibit higher values overall than black speakers and this difference can be observed in the boxplot comparisons for both the interview data as well as for the sentence data. According to the linear mixed effects analysis, there is no significant effect for ethnicity, with an increase in values for white speakers. However, there are significant effects for all other predictors. There are significant interactions between ethnicity and several of the other predictors, including f_0, F1 and RMS energy.

Based on an inspection of the boxplots, HNR values for all frequency bands are indeed lower for black speakers than for white speakers. This may indicate either added noise or voicing irregularity (Keating et al. 2015).

Incidentally, the pattern found for my data is similar to that found by Walton and Orlikoff (1994) in their study of speaker race identification in the United States, where it was found that the black speakers in their sample exhibit lower mean HNR values than the white speakers.

5.5. OVERVIEW OF RESEARCH FINDINGS

The overall patterns found for the acoustic analysis of my data as detailed in this chapter and the previous chapter, suggest that there is a difference in overall voice quality between the black and white speakers included in my study. Black speakers exhibit a lower fundamental frequency, lower HNR (and CPP) values, higher H1−H2 values (although only for the corrected measure), higher H2−H4 values, lower H4−2 kHz values and higher values for H2−H5 kHz in comparison to white speakers. The results of the statistical analysis of the interview data (the most relevant dataset for the assessment of voice quality) are provided in table 5.1 below.

Types of Measure	Measure	Wilcoxon rank-sum test results	Linear mixed effects regression results for interview data	Direction of pattern	Linear mixed effects regression results for sentence data
Measures of Spectral Slope (psychoacoustic model)	H1*−H2*	p=0.169	p=0.012*	higher for black speakers	p=0.507
	H1−H2	p=0.261	p=0.247	higher for black speakers	p=0.776
	H2*−H4*	p=0.119	p=0.121	higher for black speakers	p=0.880
	H4*−2 kHz*	p=0.011*	p<0.001***	higher for white speakers	p=0.578
	H2 kHz*−H5 kHz	p=0.627	p=0.019*	higher for black speakers	p=0.266
Other (formant-based) Measures of Spectral Slope	H1*−A1*	p=0.071	p=0.023*	higher for black speakers	p=1
	H1*−A2*	p=0.375	p=0.176	higher for black speakers	p=581
	H1*−A3*	p=0.212	p=0.781	higher for black speakers	p=0.352
Noise Measures	CPP	p=0.006**	p=0.030*	higher for white speakers	p=0.266
	HNR05	p=0.091	p=0.302	higher for white speakers	p=0.279
	HNR15	p=0.222	p=0.309	higher for white speakers	p=0.249
	HNR25	p=0.147	p=0.243	higher for white speakers	p=0.170
	HNR35	p=0.081	p=0.110	higher for white speakers	p=1
	SHR	p=0.139	p=0.084	higher for white speakers	p=0.737

Table 5.1: Statistical analysis results for the interview data.

The pattern for black speakers most closely resembles what Garellek (2016) describes as ‘unconstricted creaky voice.’ Not only does the overall pattern of the values resembles the pattern for this phonation type, for some of these measures, including H2−H5 kHz (corrected), H1−H2 (corrected) and CPP, the effect for ethnicity is significant according to the linear mixed effects analysis conducted.

Khan, Becker and Zimman (2016) refer to unconstricted creaky voice as ‘Slifka voice.’ In terms of how it is produced, they note that it has been identified using acoustic and aerodynamic studies (Slifka 2000, 2007) as involving a slowly spreading glottis with partial abduction of the ventricular folds. As for the black speakers in my study, for ‘Slifka voice’ as described by Khan, Becker and Zimman (2016) there is a lower pitch, lower HNRs (and presumably therefore also lower CPP given the high correlation between these two measures), lower SHR (as mentioned below), as well as higher values for H1*-A1*, H1*-A2* and H1*-A3*. These authors note that this phonation type may be widely used in varieties of English but underreported. Note that this phonation type (discussed in more detail in the following chapter in section 6.2.2 on page 224) would naturally entail some degree of breathiness (Keating et al. 2015) and there is therefore agreement between this hypothesis and the results that suggest greater breathiness for black speakers. In the following chapter, I discuss and ultimately argue for rejecting the 'unconstricted creaky voice' hypothesis in favour of alternative, more likely explanations for these findings (see further the discussion in section 6.2.2, page 230 in particular)

In addition to these differences, I also found significant differences (for the linear mixed effects analysis, approaching significance for the Wilcoxon rank sum tests) for the measure H1−A1 (significant for the corrected measure and approaching significance for the uncorrected measure), with the pattern of distribution indicating that black speakers have higher values for this measure in comparison to white speakers overall. This pattern suggests that black speakers may have greater posterior glottal gaps, associated with breathiness, or aspiration noise in comparison to white speakers.

The linear mixed effects analysis also revealed an effect approaching significance for SHR, with white speakers having higher values than black speakers overall for this measure. This pattern indicates that white speakers display a greater number or subharmonics to harmonics than black speakers.

The overall pattern for the H1−A2 and H1−A3 measures indicates that black speakers have higher values in comparison to white speakers for these measures. This pattern would suggest that glottal closure is more abrupt for white speakers and that black speakers may have more aspiration noise, which could be the result of non-simultaneous glottal closure.

While only some of the measures showed significant effects for ethnicity, it is not clear at this stage to what extent the observed patterns of difference are perceptually relevant to speakers of South African English and therefore whether the significant differences are more salient to South African listeners than the differences for which no significant effects were found according to the linear mixed effects analysis.

Having presented the findings of this study in this chapter and in the previous chapter and having offered some preliminary interpretation of these results, in the following chapter I will conclude with a discussion about the possible reasons for the observed differences, their potential origins and social significance, as well as directions for future research and the limitations of the current study.

CHAPTER VI: DISCUSSION, EXPLANATORY HYPOTHESES AND CONCLUSION

6.1. INTRODUCTION

In this chapter I provide a discussion of the hypotheses suggested by the research findings as presented in the previous chapter. I then discuss what is currently known about the indexical status of voice quality variation as it relates to South African English as well as the limitations of my study and directions for future research.

6.2. DISCUSSION OF HYPOTHESES BASED ON THE RESEARCH FINDINGS

In the previous chapter, I presented the findings of my study and also provided a brief discussion and suggested plausible interpretations for the findings. One of the interpretations I offered to account for the patterning of the data mentioned in the previous chapter is the possibility that the black speakers included in my sample may make greater use of a phonation type known as non-constricted creak, sometimes referred to as 'Slifka voice'. Another possibility mentioned in the previous chapter is that the black speakers may make more regular use of a phonation type such as whispery creaky voice, as defined by Laver (1980). A third possible interpretation is that while both groups of speakers may use similar amounts of creak, as suggested by the findings for the auditory analysis of the sentence data, black speakers may use a voice quality characterized by a setting which could be described as more breathy, slack or lax in comparison to white speakers, who may in contrast use a setting characterized by greater glottal stricture, stiffness and vocal fold thickness. In this section I provide a discussion of each of these possibilities with reference to the findings of the data analysis presented in the previous chapter.

6.2.2. Non-constricted Creak

In considering the first possibility, namely that non-constricted creak may be more commonly used by black speakers, it is first necessary to outline the characteristics of this phonation type in more detail. Non-constricted creak is known variously as 'Slifka voice', non-constricted voice and less constricted creaky voice (Khan, Becker and Zimman 2016). This phonation type was described by Slifka (2000, 2006) in terms of both acoustics as well as vocal tract aerodynamics, hence the term 'Slifka voice' sometimes being used to describe it.

Slifka (2000:97) describes this phonation type as one of the patterns of phonation used in the termination of utterances. Slifka (2000:104-105) observed a pattern of voicing termination for some of her subjects which did not match the pattern for either bracing of the vocal folds, or the mechanism offered by the classical descriptions of glottalization which is said to involve mostly adducted vocal folds with irregular and brief segments of voicing. Slifka's (2000:104-105) data in contrast suggests that the observed pattern involves greater abduction of the vocal folds. Slifka (2000:128-129) notes that unlike the adduction gesture specified in traditional descriptions of glottalization, the pattern observed involved irregular f_0 linked to increasing glottal area.

In a later article, Slifka (2006:171-172) points out that while for creaky phonation types such as vocal fry and (prototypical) creaky voice, strong adduction of the vocal folds is involved resulting in a decrease in airflow, some studies have found that sustained irregular vibration of the vocal folds can be maintained by increasing airflow. Slifka (2007:179) found that in many instances, for several speakers included in her study where there were f_0 irregularities associated with the termination of utterances there was also an associated increase in the mean glottal area which is consistent with abduction rather than adduction of the vocal folds towards the termination of utterances.

One of the patterns observed was found to involve abrupt closure preceding a fairly rapid glottal abduction gesture resulting in an increase in airflow for much of the cycle (Slifka 2007:183). The other pattern observed was described as involving an extensive and perhaps increasing glottal area during which there is no contact between the vocal folds (Slifka 2007:183). These two patterns were found for all of the three speakers included in Slifka's (2007:183) study who produced utterance final irregular phonation.

Keating, Garellek and Kreiman (2015), in describing different kinds of creak, also provide a description of this phonation type using the term 'nonconstricted creak.' These authors state that for nonconstricted creak, while characterized by irregular and low fundamental frequency as found for example for prototypical creak, the articulatory mechanism involved is that of a spreading rather than constricted glottis, such that there is higher rather than lower airflow through the glottis. Keating, Garellek and Kreiman (2015) note that this phonation type

has been attested as occurring in utterance-final position, with vocal fold spreading commencing before the end of the utterance. In describing the acoustic correlates of this phonation type, Keating, Garellek and Kreiman (2015) note that because there is naturally a minimal amount of subglottal pressure involved in its production and because in addition the glottis is spreading, the conditions for sustaining voicing are unfavourable. Keating, Garellek and Kreiman (2015) point out that while because of the increase in airflow for this phonation type, there is consequentially some degree of breathiness involved, it is nevertheless distinct from the phonation type proposed by Laver (1980), said to involve airflow though the arytenoid gap combined with creak produced anteriorly[55]. It is therefore important in interpreting the results of the current study at least in as far as they can be accounted for in terms of differences in creak to consider whether there is more convincing evidence in favour of the greater use of non-constricted creak on the part of the black speakers included in the study or whether there is any evidence suggestive of airflow through the arytenoid gap.

Khan, Becker and Zimman (2016) discuss several important points regarding non-constricted creak. In their study, social and phonological variation was controlled for by restricting the analysis to IP-final words from the speech of five transgendered men who were speakers of American English. As noted by Zimman (2013), this demographic is associated with a greater use of creak among American English speakers.

Khan, Becker and Zimman (2016) found a strong correlation between higher ratings of creak and low f_0. However, Khan, Becker and Zimman (2016) also found an unexpected effect, although non-significant, namely higher values for H1*−H2*, H1*−A1*, H1*−A2* and H1*−A3* associated with higher ratings of creak. Lower HNR values were found to be correlated with higher ratings of creak as would be expected (Khan, Becker and Zimman 2016). In addition and somewhat unexpectedly, lower values were observed for SHR. Khan, Becker and Zimman (2016) suggest that this pattern could best be accounted for as a result of their speakers using 'Slifka voice' as opposed to other kinds of creaky phonation.

[55] In terms of the auditorily identified phonation types mentioned earlier, the phonation type described by Laver (1980) here is probably closest to whispery creak. This, however, is only an approximate equivalence given that whispery creak is operationalized in this study as a phonation type involving creak with evidence of accompanying aspiration (i.e. operationalized in acoustic terms), while the phonation type described by Laver (1980) is defined in primarily articulatory terms. As this is not an articulatory study and as such, we do not have evidence of exactly which articulators are involved in producing this auditory impression (and accompanying acoustic markers), to directly equate whispery creak with the voice quality described by Laver (1980) would be imprecise at best. The whispery creaky voice described by Laver (1980) is in any case dubious (Keating p.c.).

As noted in the previous chapter, the patterns for the acoustic measures observed in my data for black speakers in comparison to white speakers appear to resemble the findings of Khan, Becker and Zimman (2016). That is, black speakers display a trend towards lower SHR values, higher values for H1*−H2*, H1*−A1*, H1*−A2* and H1*−A3* and lower values for f_0, HNR and CPP when compared to white speakers. Given this resemblance, it would seem reasonable to suggest that the differences between black and white speakers could potentially be explained by an account in which black speakers use non-constricted creak to a greater extent than white speakers. However there are several reasons why this interpretation may not be the most likely one given other alternatives.

Firstly, it should be noted that Khan, Becker and Zimman's (2016) findings pertain to IP-final words only. The use of non-constricted creak as noted by Keating, Garellek and Kreiman (2015) is also only attested utterance-finally. The VS measurements on which the main findings of my study are based were however taken for all vowels regardless of whether they occurred utterance finally or not, in line with the research focus on overall voice quality differences. Were non-constricted creak used utterance finally in the variety of South African English under investigation, it is not clear whether such well-defined and consistent patterns would have resulted from an analysis of measurements taken from vowels occurring in all utterance positions within a semi-structured interview.

Secondly and of even greater relevance, Khan, Becker and Zimman (2016) focused specifically on those utterances which were given higher ratings for creak. That is, Khan, Becker and Zimman (2016) were able to directly evaluate the acoustic characteristics of one phonation type, namely creak, against what would be expected given the traditional descriptions of this phonation type. However, my study is concerned not with only one phonation type, but overall voice quality thereby including several different phonation types which may be present in the interview data. If I were comparing only those vowel segments during which only creaky phonation types were present, the claim that the differences for the acoustic measures arise from greater use of non-constricted creak on the part of black speakers would have more weight. This not being the case however, there is little additional evidence in support of the hypothesis that

the differences observed between black and white speakers have more to do with differences in the use of non-constricted creak.

Keating, Garellek and Kreiman (2015) have provided an example waveform illustrating non-constricted creak. This image is reproduced below in figure 6.1.

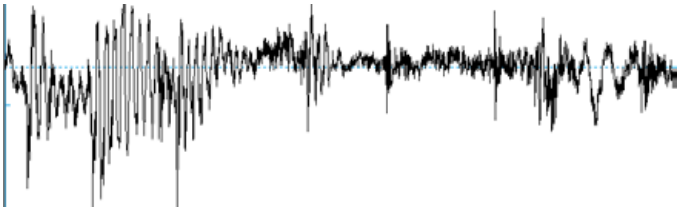

Figure 6.1: Example of nonconstricted creak phrase-finally for a male English speaker as provided in Keating, Garellek and Kreiman (2015).

By way of comparison with my data, I have also provided an example of the waveform display illustrating the appearance of the closest example to non-constricted creak as described by Keating, Garellek and Kreiman (2015) selected from the data of black speakers who most clearly exhibit this pattern. It is likely that if there is evidence of non-constricted creak in the sentence data, it would more likely be found for the phonation type described in chapter three as whispery creak, where some aspiration noise is anticipated as well as some degree of waveform irregularity. The waveform of one of the examples which provides the closest comparison to non-constricted creak in this category for the sentence data is displayed in figure 6.2 below.

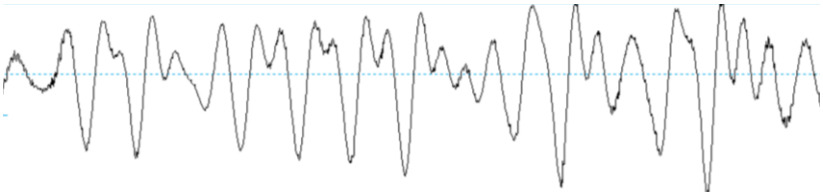

Figure 6.2: A waveform display of a vowel extracted from the token *he* from the sentence data of speaker S3.

While there is some irregularity present as well as some signs of aspiration noise, the pattern displayed in figure 6.2 above, does not appear to provide any clear suggestion of increasing glottal abduction to the extent that the example of non-constricted creak provided in figure 6.1[56] does, although as Garellek (2016) cautions, waveform displays alone may not always be particularly useful in discriminating between tokens containing greater or lesser degrees of breathiness.

This example is nevertheless quite possibly the closest pattern to that of non-constricted creak present in the sentence data of this study. Other examples of the phonation type 'whispery creak' are similar. It is of course anticipated that given that the sentence data were derived from measurements taken from utterance medial prominent syllables that a phonation type such as non-constricted creak would not be likely to be used in this data set in any case, since it has been attested only utterance finally in previous studies. Therefore it could be the case that for the interview data, derived from measurements taken at all utterance positions and with different degrees of prominence, the use of such a phonation type would be more likely to occur and that the acoustic consequences of such differences in use between the speakers of the two ethnic groups are what are reflected in the patterns of difference between black and white speakers observed in this study.

Given this possibility, I examined the interview data, selecting speakers whose voice profile indicated average values for those measures most consistent with the description of non-constricted creak as discussed above, for example, high values for spectral tilt measures and lower values for noise measures. After having identified several such speakers, I listened and searched through the data for possible examples of non-constricted creak occurring in these interviews. Having found several potential examples of these, I subjected the samples to an acoustic analysis to test whether in particular, along with the other signs of non-constricted creak, values for H1*−H2* were particularly high rather than low or negative as this is one of the key diagnostic criteria used in identifying non-constricted creak (Keating, p.c.) and distinguishing it from other types of creak. Waveforms of the potential examples of non-

[56]Unfortunately, the time scale for figure 6.1 is not known (and therefore only a roughly equivalent comparison can be made rather than a direct comparison). Figure 6.1 represents a phrase final utterance.

constricted creak used by these speakers are presented in figures 6.3 and 6.4 below. None of these samples displayed particularly high H1*−H2* values, being close to 0 in each case and are therefore unlikely to be examples of non-constricted creak (Keating p.c.). Thus the relatively few examples which could potentially have been non-constricted creak even for the interview data, for those speakers who exhibit an overall vocal profile most consistent with the acoustic characteristics of this phonation type are in fact not likely to be non-constricted creak examples.

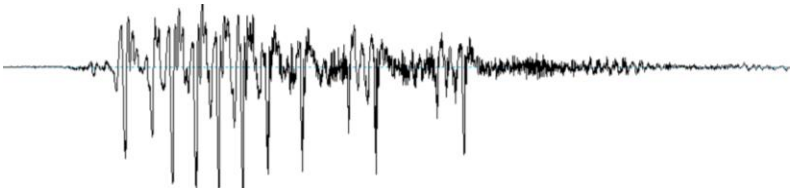

Figure 6.3: Waveform display of the word *ja* as spoken by speaker B2.

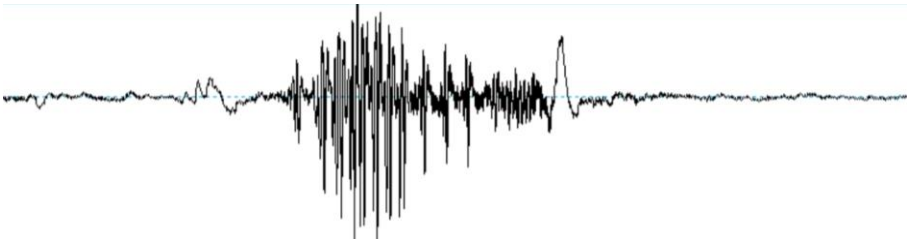

Figure 6.4: Waveform display of the word *ja* as spoken by speaker L2.

Given therefore that there is minimal evidence of the use of non-constricted creak in my sample generally, even for those speakers whose overall vocal profile most closely matches the pattern expected for non-constricted creak, it is necessary to consider other possible explanations for the observed patterns of difference which may perhaps be more plausible. One of these which I have already suggested in the previous chapter is that black speakers may make more regular use of a breathy, slack or lax voice quality to a greater extent in comparison to white speakers

who may make more habitual use of a voice quality involving greater stiffness and/or tension which could potentially be described as 'pressed'. I will consider and evaluate the evidence with regards to this hypothesis in the following section.

6.2.3. The Breathy/lax/slack Versus Pressed/tense/stiff/constricted Voice Hypothesis

Laver (1980:139) describes a compound phonation type involving three components, namely voice, creak and whisper, thus yielding 'whispery creaky voice'. For this phonation type, Laver (1980:139) speculates that in terms of production it would involve creak produced at the thyroid portion of the glottis, such that the cartilaginous part of the glottis can be used for producing whisper and so that the ligamental glottis can be used to provide the voicing component. It is possible that the examples provided in figures 6.3 and 6.4 above are more likely to be examples of such a whispery creaky voice rather than non-constricted creak.

This phonation type could arguably occur in cases where for example, creak is used in similar proportions by both groups but breathiness (particularly if aspiration noise is produced at the cartilaginous portion of the glottis) is used more by one of the groups and where such breathiness occurs simultaneously with creak. My findings for the sentence data, although not definitive, generally provide some support for the hypothesis that the black speakers use a type of breathy voice with somewhat greater frequency than the white speakers. The auditory analysis of the sentence data revealed that black speakers use breathy voice as auditorily identified with greater frequency than white speakers do and also make greater use of whispery creak as auditorily identified, although in both cases by a relatively slight margin. It is therefore important to consider whether the observed differences are likely to be a result of greater use of such a phonation type by the black speakers as would arise from having similar levels of creak to white speakers but with the addition of some form of aspiration noise such as that typically found for breathy voice.

As noted by Fung (2015), for breathy voice, CPP is expected to be relatively small, while for the other spectral parameters such as H1*−H2*, H2*−H4*, H1*−A1*, H1*−A2* and H1*−A3*, breathy voice should have comparatively larger values. This pattern is indeed that

observed for the black speakers in comparison to the white speakers in my study, consistent with the hypothesis that black speakers make use of a more breathy voice quality, while white speakers make use of a tenser, or more pressed voice quality overall.

As noted by Kuang and Lieberman (2015), a breathier voice is associated with a steeper spectral slope (as reflected in the values for H1*−A3* and H1*−A2* for example), while a flatter spectral slope has an association with either a tenser or a creakier voice. The values for these measures in my study show a trend towards white speakers having lower values for these spectral slope measures, indicating a flatter spectral slope for white speakers, suggesting that they may make use of a tenser voice quality in comparison to black speakers. That white speakers may, in general, make use of such a setting is also suggested by the higher f_0 observed for white speakers (for both the sentence and interview data), since there is a natural association between the use of a higher f_0 and the production of tense voice, according to Kuang and Lieberman (2015) citing Titze (1988).

Conversely, as noted by Laver (1980:132), it is quite rare to encounter high-pitched breathy voices. Gordon and Ladefoged (2001:398) have also noted the consistent association between tone-lowering and breathy phonation cross-linguistically. With falling f_0 as found by Hirano, Ohala an Vennard (1969), cited in Garellek (2013:125), there is evidence of relaxation of the thyroatrytenoid and cricoarytenoid muscles, which could conceivably produce a more breathy voice quality. As Klatt and Klatt (1990: 824) note, Pandit (1957) had also found that there was a tendency for breathy vowels to be lower in fundamental frequency, which Klatt and Klatt (1990) attribute to the need to slacken the vocal folds in order to maintain voicing despite the separation at the posterior end of the folds. The fact that fundamental frequency is consistently lower for black speakers than for white speakers across contextual styles is once again in accordance with the hypothesis that a breathier mode of phonation characterizes the voice quality of black speakers as a group and a mode of phonation involving greater tension characterizes the voice quality used by white speakers overall.

The results for the spectral tilt measures, especially when considered in combination with the results for the noise measures (including CPP and HNR) would favour the interpretation that there is greater use of some form of breathy phonation on the part of black speakers.

Labuschagne and Ciocca (2016:198-198) point out that CPP has shown a consistently strong negative correlation with breathiness in a number of studies, as there is a correspondence between lower periodicity and greater breathiness and for this reason it is one of the strongest acoustic correlates of breathiness. As noted by Garellek and Keating (2011:187), CPP has been used to successfully distinguish between breathy versus non-breathy voice qualities in terms of both perception and production, citing work by Blankenship (2002) and Esposito (2006 , 2010).

Given the importance of CPP as an acoustic measure closely correlated with perceptual breathiness, it is worth pointing out that the effect for CPP is perhaps the most consistently observed effect in my study. I observed the same pattern for both the sentence data as well as the interview data.Black speakers were consistently found to have lower values overall for this measure in comparison to white speakers.

In their study, Garellek and Keating (2011:194) found that for men, values for H1*−A2* were higher, but values for CPP lower, concluding from this finding that the men in their sample may be breathier. In my sample, this is a consistent pattern for black speakers across both the interview data and the sentence data, suggesting similarly that black speakers may be breathier. Garellek and Keating (2011:194) note that while lower CPP values alone could be attributed to greater laryngealisation for men than for women, when considered in conjunction with the higher H1*−A2* values observed for the men in their study, that men may be breathier is a more likely conclusion. They also note that there is a corresponding difference for H1*−A1* values, with higher values for this measure for men (Garellek and Keating 2011:194). In my study, I found a similar pattern in terms of the acoustic measurements to that of Garellek and Keating (2011), with black speakers having higher values for H1*−A2* and H1*−A1* but lower values for CPP in comparison to white speakers, lending further support to the hypothesis that black speakers in general habitually make use of a breathier phonation type when compared to white speakers. If such patterns are indeed as a result of a difference in terms of breathiness, it is worth considering how such a difference may be achieved in articulatory terms based on the observed patterns.

As Blankenship (1997:44) notes, there are three main mechanisms which may be used in the production of breathy voice. As already mentioned, the duration of vocal fold closure may be reduced, which may possibly result in higher values for H1*−H2*, vocal fold closure may be

more gradual, which may possibly result in a steeper spectral slope (reflected for example, in H1*−A2*) and finally, there may be a posterior glottal opening or 'chink' present, allowing air leakage as well as frication through the glottis (reflected in a widening of the first formant bandwidth and an increase in H1*−A1*). Importantly, Blankenship (1997:44) points out that of this suite of articulatory gestures, certain gestures may be favoured by speakers of certain languages over others in creating contrasts in terms of phonation. Thus open quotient may be more important for some languages and therefore dialects (Ní Chasaide and Gobl 2010) than for example, the use of aspiration noise (Blankenship 1997:44). Esposito (2010) also found that certain parameters were more important than others in forming the phonological distinction between breathy and other phonation types for certain languages. There is some evidence that all of these ways of realizing a more breathy/lax/slack phonation type among the black speakers of my study are used to different degrees, with some possibly being more important for some speakers than others.

Below I discuss the evidence indicating which mechanisms appear to be most commonly and consistently used. The findings for the noise measures may suggest that overall, the most consistent difference is that of greater aspiration noise for black speakers, which could presumably be attributed to incomplete glottal closure as a result of weaker activation of the lateral cricoarytenoid, interarytenoid and thyroarytenoid muscles (Zhang 2016b:2629). As pointed out by Zhang (2016b:2629), incomplete glottal closure can be located either at the cartilaginous portion of the glottis or the membranous (ligamental) portion. There is some evidence that some of the speakers in my study rely more on one than the other to produce incomplete glottal closure and therefore aspiration noise. These two main options are discussed in more detail below as they relate to hypothesized differences in the abruptness of vocal fold closure, nonsimultaneous glottal closure, differences in open quotient and glottal stricture as well as differences in the size of the posterior glottal gap.

6.2.4. Hypothesized Differences in the Abruptness of Vocal Fold Closure

There is some evidence to suggest that black and white speakers differ in terms of overall trends in the abruptness of vocal fold closure. There is a relatively consistent trend across contextual styles for black speakers to exhibit somewhat higher values for H1*−A2*. Higher values for this measure were interpreted by Blankenship (1997) as resulting from a less abrupt closing gesture,

which may suggest that in my study, the closing gesture for black speakers is somewhat more gradual than for white speakers.

The mechanism whereby this difference may be achieved is that of a relaxation of the vocal folds sufficient in order for them to meet in a more wavelike fashion (Blankenship 1997:20) on the part of black speakers and a tenser configuration being used by white speakers. This would again be consistent with a hypothesis of a breathier voice quality being used by black speakers and a less breathy, tenser mode of phonation being used by white speakers.

Both of the measures H1*−A2* as well as H1*−A3* are known to show some correlation with spectral tilt overall, which may be due to the abruptness of glottal closure (Stevens 1977 and Hanson et al 2001 cited in Garellek 2012:7). In my study, for these measures I found that for both, white speakers have lower values overall in comparison to black speakers, suggesting that white speakers have flatter spectral tilts, suggesting in turn that white speakers make use of a more abrupt glottal closing gesture than black speakers. This once again supports the notion that a more breathy voice quality is used by black speakers in comparison to white speakers.

6.2.5. Hypothesized Nonsimultaneous Glottal Closure

Another means of achieving a breathier voice quality is by nonsimultaneous closure of the glottis and there is evidence from my findings which point towards a general trend for such nonsimultaneous closure to be more prevalent among black speakers than white speakers. As discussed by Stevens and Hanson (1995:150), many speakers use a vocal fold configuration that ensures that the glottis never reaches full adduction throughout the glottal vibratory cycle. One way of achieving this, following Stevens and Hanson (1995:150) is to maintain abduction of the vocal processes for the duration of the glottal cycle. The acoustic effect would be to increase spectral tilt, particularly reflected in values for H1*−A3* such that there are higher values for this measure. In the findings reported in the previous chapter it is clear that overall, black speakers have higher values for this measure than white speakers do.

However, while these results do suggest that there may be some differences in terms of nonsimultaneous glottal closure between black and white speakers as a general trend, there are

reasons to believe this is not necessarily the primary mechanism whereby a more breathy type of voice quality may be achieved.

In my study the results of the analysis revealed a difference of approximately 1.9 dB for H1*−A3*. However, for H1*−A3, Hanson and Chuang (1999:1076) observed a 9.6 dB difference between males and females, which is considerably higher than the difference observed between black and white speakers in my study. Thus comparatively, the differences overall are considerably less than between males and females (at least based on Hanson and Chuang 2001:1075) for H1*−A3*.

Hanson et al (2001: 462-463) divided their speakers into two groups based on whether they had H1*−A3* values higher than 23 dB and those lower than 23 dB, with the former being labeled group two and with that group hypothesized to use nonsimultaneous glottal closure during phonation. Very few of my speakers have mean H1*−A3* values of 23 dB and above. In fact there are only three of them, two of whom are white. This may suggest, based on the smaller than expected difference in terms of H1*−A3* for my speakers, that nonsimultaneous glottal closure is not the most commonly used way of producing the overall difference in terms of voice quality between black and white speakers and that other means are more salient.

6.2.6. Hypothesized Differences in the Size of the Posterior Glottal Opening

Another frequently mentioned mechanism for producing more breathy phonation is that of maintaining a posterior glottal gap, or 'chink' during phonation. Following Hanson et al (2001), cited by Garellek (2012:7), the H1*−A1* measure shows a correlation with breathiness and is thought to relate to the presence of a posterior glottal gap specifically (located at the cartilaginous glottis). Klatt and Klatt (1990:852) describe some of the vocal tract aerodynamics which may be used to effect some of the acoustic differences observed for breathy voice. As a result of a posterior glottal chink, there will be an increase in losses at low frequencies in the transfer function of the vocal tract, thus increasing first formant bandwidth (Klatt and Klatt 1990:852). They speculate that what happens in this case in the production of breathy voice is the inward rotation of the arytenoid cartilages in order to compensate for the presence of a posterior glottal gap as well as a slackened posture of the vocal folds necessary to produce an appropriately low fundamental frequency (Klatt and Klatt 1990:852). This would of course also

be consistent with my findings for black speakers who exhibit a lower fundamental frequency than white speakers overall.

Stevens and Hanson (1995:150) mention that an increase in first formant bandwidth (and therefore H1*−A1*) is expected as a result of approximating the vocal processes posteriorly while maintaining a fixed gap between the arytenoids. In addition, there will also be some increase in tilt (reflected by the measure H1*−A3*) and there will also be high frequency turbulence noise (acoustically typically evaluated by means of noise measures).

Perhaps one of the clearest differences between the black and white speakers in my sample is the difference for H1*−A1* which both in terms of magnitude, appearance in terms of the graphical comparison and in terms of the statistical analyses, most consistently pertains both to the interview as well as to the sentence data between the two groups of speakers. The white speakers overall displayed lower values for this measure in comparison to the black speakers, suggesting that white speakers may have smaller posterior glottal gaps. Such differences could also account to some extent for the higher values for the other spectral tilt parameter H1*−A2* observed for black speakers. Maintaining a posterior glottal gap can be achieved by means of thyroarytenoid muscle activation without attendant activation of the lateral cricoarytenoid and interarytenoid muscles (Zhang 2016b:2617).

However, while larger posterior glottal openings may be more prevalent among black speakers than among white speakers overall, there is some reason to consider that the actual differences might be relatively small compared to what has been observed in some other studies. For H1*−A1* based on my interview data, a range of 8.55 dB was observed for black speakers and a range of 8.38 dB for white speakers, which are considerably less than the range of 16 dB observed for Stevens and Hanson's (1995:161) speakers. This may suggest a more restricted range of glottal opening sizes in my South African sample, with the range in terms of glottal gaps being slightly greater for black speakers.

Finally, while there is evidence to suggest larger posterior glottal openings for the black speakers in my study in comparison to white speakers, this should not automatically be assumed to be the main articulatory correlate of the hypothesized difference in breathiness between the two groups in perceptual terms, since as noted by Samlan et al (2013:1210), there is not always a

high correlation between breathiness and the size of the glottal gap, citing Rammage, Peppard and Bless (1992) and Södersten and Lindestad (1990).

6.2.7. Hypothesized Differences Relating to Open Quotient and Glottal Stricture

In addition to evidence of larger glottal openings being used by black speakers as well as nonsimultaneous glottal closure in comparison to white speakers, I have also found evidence suggestive of some differences in terms of open quotient (OQ) and glottal stricture. I found a trend towards higher values for H1*−H2* for black speakers, suggesting greater OQ, although this was only clearly observable for the interview data. Thus it seems that while greater OQ may be used for some speakers and does clearly play a role in producing breathier phonation, it is perhaps more common across contextual styles for speakers to use a form of incomplete glottal closure. This could suggest that black speakers adjust OQ more to achieve less breathiness in more formal speaking styles, or that white speakers adjust OQ to achieve greater breathiness in such contextual styles. Because the sentence data also consisted of the analysis of prominent syllables only, the differences across these two data sets may also therefore indicate a different approach to the realization of OQ as a type of prominence effect.

As discussed by Chen, Park, Kreiman and Alwan (2014), H1*−H2* in addition to being a reflection of OQ, is also associated with glottal pulse skewness, quantified by SQ or speed quotient, although this may not always have a monotonic relationship with SQ and may also vary independently as a function of f_0. Given that there is based on the other measures, some indication that overall black speakers may make use of larger posterior glottal openings than white speakers, it is not necessarily the case that the differences observed reflect differences in open quotient, but rather differences in glottal stricture (Chen, Park, Kreiman and Alwan 2014), suggesting that the glottal setting for white speakers is generally more constricted than that for black speakers. More recently, Zhang (2016a) found that the effect of increasing the thickness of the medial section of the vocal folds reduces H1−H2, which may also suggest that to some extent white speakers may also exhibit greater medial vocal fold thickness in comparison to black speakers in general.

6.2.8. Other Acoustic Evidence Supporting the Breathy/slack/lax Versus Pressed/stiff/tense/constricted Hypothesis

The possibility that larynx lowering may be involved in both the realization of some breathy voice for black speakers as well as the observed lower fundamental frequency receives some support from a comparison of the values for the third formant, which are higher overall for white speakers in comparison to black speakers, suggesting that there is either larynx raising on the part of white speakers or that there is larynx lowering on the part of black speakers or possibly some combination of the two. Klatt and Klatt (1990:825) drew a similar conclusion regarding the acoustic effect of larynx raising on F3. Larynx lowering would be consistent with a generally more relaxed setting.

While the observed higher SHR values for black speakers for the interview data are potentially more difficult to account for in terms of a difference in breathiness, as discussed in Švec, Schutte and Miller (1996) it is possible for an increase in subharmonics to be linked to modes of phonation involving reduced adduction and increased airflow, consistent with a breathy type of phonation. While resembling vocal fry in terms of subharmonics, such modes of phonation can be distinguished from vocal fry by both the greater OQ values as well as by higher airflow rates (which are also higher than modal phonation) (Švec, Schutte and Miller 1996). Such a pattern can result from both an asymmetrical closing gesture as well as certain patterns of lateral upper movement of the vocal folds interfering with their closing (Švec, Schutte and Miller 1996).

Further support for the hypothesis that black speakers make use of a more breathy voice quality can be found in the comparison of the data for measurement points coded as modal in the sentence data analysis. Given the auditory coding procedure as described in the methodology chapter, we would expect that if there is any difference within the modal category, it would be in terms of breathiness rather than creak. Within the modal category, there are distinct differences between black and white speakers for measures such as CPP and H1*−A2* as can be seen from figures 6.5 and 6.6 provided below. That such patterns should persist within the modal category suggests that rather than the difference being a result of the habitual use of certain types of creak (which would also result in lower levels of HNR and CPP for black speakers if modulation noise were present and some creak types such as non-constricted can also mimic phonation types involving some degree of whisper), it is more likely that the consistent patterns observed can be

accounted for in terms of a difference in breathy/slack/lax voice quality being used by black speakers versus a tenser voice quality being used by white speakers.

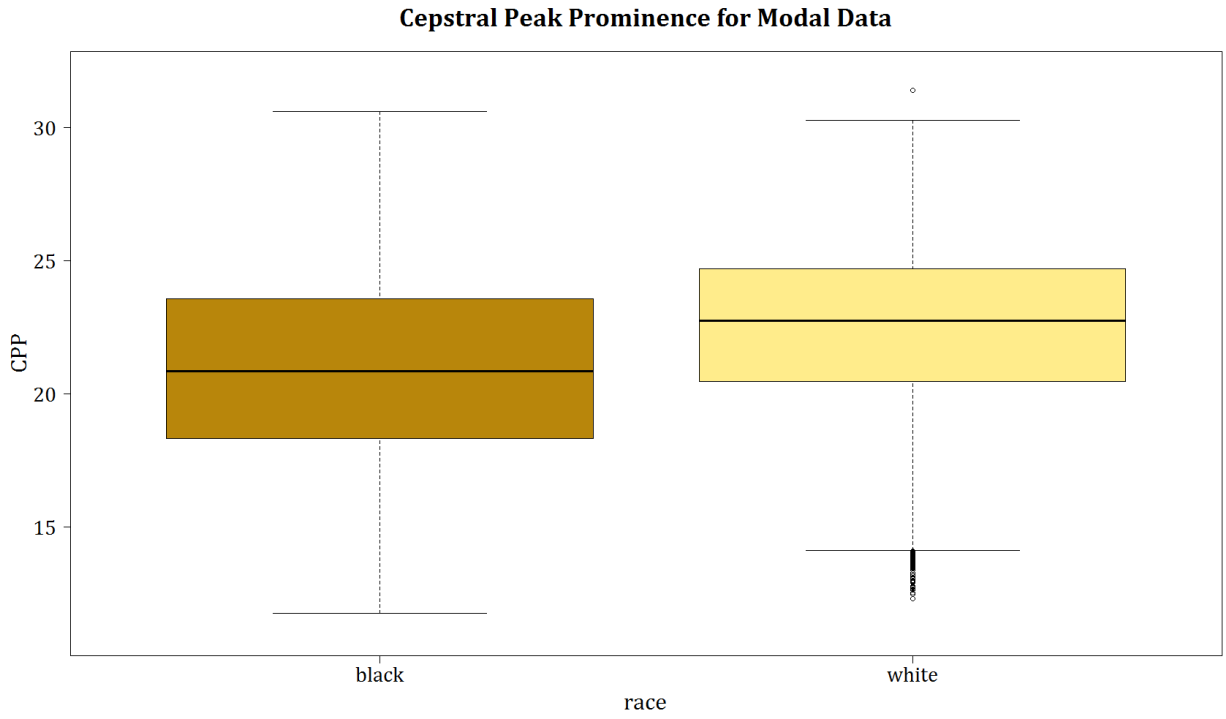

Figure 6.5: Boxplots representing the data distributions around the median for black and white speakers for CPP data for all data measurement points coded as modal.

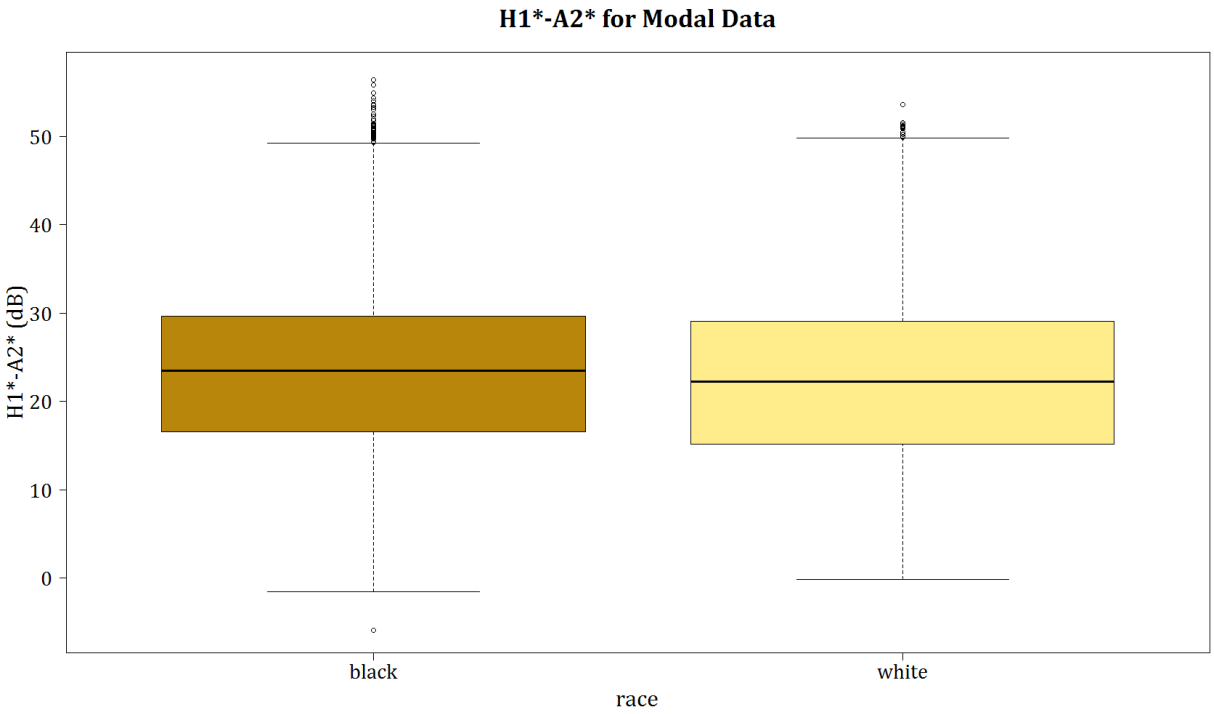

Figure 6.6: Boxplots representing the data distributions around the median for black and white speakers for H1*−A2* data for all data measurement points coded as modal.

Further evidence supporting the notion that white speakers make use of a setting involving greater vocal fold stiffness is supplied by the values for H2*−H4* where black speakers have higher observed values overall. According to Garellek (2012:7), citing Zhang et al (2011), the measure H2*−H4* is considered to have a correlation with stiffness of the vocal folds (defined as a property of the folds, representing the "elastic restoring force in response to deformation" (Zhang 2016b:2626) and may also be used in breathiness perception (Kreiman et al. 2011). Zhang et al (2013:457) also found that there was a significant correlation between H1−H2 and vocal fold body stiffness with increasing body stiffness reducing H1−H2 and additionally lowering noise production (Zhang 2016a:1501), both characteristics of the overall pattern for white speakers included in my study. According to Zhang (2015), anterior-posterior stiffness of the vocal folds is a main determinant of fundamental frequency, increasing with increasing stiffness. Higher f_0 is, as already mentioned, observed as a consistent pattern for white speakers in comparison to black speakers overall in my study.

Differences in noise production (as reflected in values for example HNR and CPP) are some of the most consistent differences between black and white speakers included in my study.

Decreases in noise production are associated with increased anterior-posterior stiffness, decreases in resting glottal angle and decreases in subglottal pressure (Zhang 2016a:1505). In comparison to black speakers, white speakers show evidence of decreased noise production on the whole (as reflected in higher CPP and HNR values) which may therefore suggest that white speakers make use of decreased subglottal pressure (also suggested by the lower RMS energy values observed for white speakers), a decreased resting glottal angle (consistent with for example, the decreased posterior glottal gap as mentioned earlier relative to black speakers) and increased stiffness. As noted by Zhang et al (2013:461), citing Hirano (1974), vocal fold body stiffness is primarily regulated by the thyroarytenoid (TA) muscle suggesting that potentially TA muscle activation plays an important role in contributing towards the overall differences observed between black and white speakers in my study. That vocal fold stiffness is not independent from initial glottal width (Zhang 2015:907) also matches the results for the formant amplitude measures in my study, which would suggest that glottal width may be greater for black speakers. The results therefore point to white speakers having a setting which is characterized by more stiffness (potentially as a result of greater activation of the TA muscle) and presumably less breathiness than black speakers.

6.2.9. Interpretations Based on More Recent Models

The foregoing analysis of the findings in terms of the breathy/slack/lax versus pressed/tense/stiff voice hypothesis relies to a large extent on measures of spectral tilt such as H1*−A1* and H1*−A3*, that is, measures of spectral tilt which are based on formant frequencies. While these measures are expected to correlate well with perceptually relevant differences in voice quality, given that they have been found to be phonologically important in several languages which have phonemic voice quality distinctions (see in particular, Keating and Esposito 2006 and Esposito 2010), there are other spectral tilt measures which may also be perceptually relevant as suggested by certain models. One such model is Kreiman, Gerratt, Garellek, Samlan and Zhang's (2014) psychoacoustic model (as described in chapter two) which represents spectral tilt based on frequency band parameters. One advantage of modeling spectral tilt based on frequency band parameters as opposed to formants is that formant-based parameters will often overlap with the former in any case, due to the fact that in a source model, the amplitudes of formants are not defined (Garellek 2016:17), whereas such overlap does not occur with frequency-delimited measures. Secondly, using measures such as H1−A1 requires the assumption that the manner in

which voice quality relates to spectral tilt is to some extent dependent on vowel quality, whereas this assumption is not as essential for a model which relies on frequency-based measures (although as pointed out by Garellek (2016:17), Kreiman et al. (2014) nevertheless assume that voice quality can be influenced by both the source as well as the filter).

The parameters used to model spectral slope in Kreiman et al's (2014) psychoacoustic model are H1−H2, H2−H4, H2−4 kHz and H2−5 kHz. In general, for most of these measures, my data reveal patterns of difference which are consistent with the general patterns described above, mostly with reference to the formant-based measures. Thus for example the difference for H1*−H2* is that black speakers have higher values overall for this measure in comparison to white speakers and this difference also happens to be significant in terms of the linear mixed effects regression model for the effect of ethnicity. For H2*−H4* there is also likewise a difference which shows that black speakers have in general a steeper spectral tilt than white speakers do. For 2 kHz*−5 kHz, black speakers also showed a significantly greater spectral tilt than white speakers. The measure H4*−2 kHz* however, did not show a very clear or obvious difference between the two groups, although the results of the linear mixed models regression analysis indicated a significant effect in terms of white speakers exhibiting greater spectral tilt for this measure. As mentioned in the results chapter in discussing these results however, this could potentially be accounted for as a purely ancillary effect of the pattern for the other spectral tilt measures in the psychoacoustic model (Garellek et al. 2016). Thus the results for the majority of measures point towards an overall difference between the two groups in terms of black speakers exhibiting greater spectral tilt than white speakers for most frequency bands which would be consistent with the findings for the formant-based measures.

As stated by Garellek (2016:16), two main dimensions define voice quality, namely noise and spreading versus constriction and both are related to specific acoustic characteristics in terms of Kreiman et al's (2014) psychoacoustic model. Garellek (2016:16) notes that the correlate of increased constriction or alternatively, increased spreading of the vocal folds which has been found to be most reliable is that of increasing spectral tilt associated with increased spreading and with the opposite applying for increased constriction. In Kreiman et al's (2014) psychoacoustic model, the increase in spectral tilt associated with an increase in vocal fold spreading as would be the case with breathy voice would be reflected in higher values for the

parameters of H1−H2, H2−H4, H4−2 kHz and H2 kHz− H5 kHz, the first two of these being the most phonologically relevant (Garellek 2016:16). In my study, for most of these measures, particularly for the first two of these, there is clear evidence of a trend towards black speakers having greater spectral tilt than white speakers, suggesting either increased spreading of the vocal folds (as would be the case with breathy voice) for black speakers or increased constriction for white speakers or some combination of these overall.

It is important to note that some of the most robust differences in terms of the spectral tilt measures included as part of Kreiman et al's (2014) psychoacoustic model were found for H1*−H2* and H2*−H4* in my study as discussed above. This is worth pointing out because several studies have found that to some degree, listeners are sensitive to distinctions in terms of H1−H2 in distinguishing between modal and non-modal voice qualities (Garellek 2016:16, citing Kreiman et al. 2010 and Kreiman and Gerratt 2010). Bishop and Keating (2012), cited in Garellek (2016:16) have also found that the parameter H2−H4 has some relevance for speaker sex identification and so therefore is likely to be of some perceptual relevance. Thus, to the extent that the findings for these measures match the interpretation of the findings for the other measures examined in this study, they provide additional evidence of the breathy/slack/lax versus constricted/tense/pressed hypothesis put forward in this chapter.

However, as noted by Garellek (2016:16), it is not clear yet whether the spectral slopes at higher frequencies included in this model, namely, H2 kHz− H5 kHz and H4−H2 kHz contribute in any relevant way towards differences in terms of linguistic voice quality. Thus, even though there is a statistically significant difference (according to the regression models used) for the H4−H2 kHz measure and a significant difference for the H2 kHz−5 kHz measure between the two ethnolinguistic groups in this study, it is uncertain at this point whether this equates with any particular perceptual difference. However, as Garellek (2016:16) points out, these two measures in addition to the other measures of spectral tilt included in Kreiman et al's (2014) psychoacoustic model may play some role in differentiating creaky from non-creaky vowels at least for American English, citing research by Garellek and Seyfarth (2016). Given that at least for the sentence data, the use of prototypical creak was found to be quite similar in terms of frequency for both black and white speakers, it is possible that similar levels of creak may have

had an obscuring effect for these measures. Even so, clearly more investigation into the perceptual relevance of these acoustic measures (particularly for H4−H2K, given the highly significant difference for this measure and that, contrary to other measures, white speakers exhibit higher values) for South African listeners is needed.

As noted by Garellek (2016:24), Zhang (2016b) has recently advanced a model systematically relating acoustic parameters, some of which feature as part of Kreiman et al's (2014) psychoacoustic model, to the articulatory parameters. For this reason, Garellek (2016:24) considers Zhang's (2016a) model to be helpful in making an assessment of the causal relationship between perception, acoustics as well as articulation and I therefore provide an interpretation of my research findings based on the data according to Zhang's (2016a) model below.

One of Zhang's (2016a) findings is that there are three parameters which the rate of vocal fold vibration depends on (Garellek 2016:27). These include subglottal pressure, stiffness of the vocal folds and approximation of the vocal folds (Garellek 2016:27). Zhang (2016a) also found that, regardless of subglottal pressure it is possible to raise f_0 by means of a glottal width decrease, although this will usually be concomitant with greater thickness of the vocal folds and decreased stiffness of the vocal folds along the front-to-back dimension. This would result in both a higher f_0, greater constriction and as a result a decrease in spectral tilt, as measured by the parameters used in Kreiman et al.'s (2014) psychoacoustic model (Garellek 2016:27). These, as Garellek (2016:27) points out, are features characteristic of tense voice and this accords well with the interpretation provided in this chapter for white speakers, who overall display higher fundamental frequency and decreased spectral tilt suggesting a tenser type of voice quality than black speakers. According to Zhang's (2016a) model, a reasonable interpretation of my findings would be that white speakers may use such mechanisms as decreasing glottal width (in comparison to black speakers who show evidence of increasing width for example), possibly by decreasing the size of the posterior glottal opening with the associated increase in vocal fold thickness and decrease in vocal fold stiffness on the front-to-back dimension in order to achieve this effect.

As Garellek (2016:28) notes, the parameter of 'vocal fold thickness' has a strong impact on spectral tilt, such that measures such as those included in Kreiman et al's (2014) model would

be lower for increased thickness of the vocal folds. Thus it is plausible that the white speakers in my study generally may make use of increased vocal fold thickness to an extent that black speakers do not. Because following Zhang's (2016) model, the thickness of the vocal folds has a greater impact than glottal width in the production of voice qualities associated with certain spectral tilt changes, it may be that, at least according to this model, the impact that the hypothesized increased vocal fold thickness for white speakers may have is more important in terms of voice quality perception than the hypothesized gestures of maintaining a glottal gap and non-simultaneous glottal closure hypothesized for black speakers. As Garellek (2016:28) cautions however, it is likely that in actual human voices, there will be some covariation between vocal fold thickness and other model parameters in ways not yet fully understood and therefore, in future, as our understanding of the perceptual significance of these different parameters increases, it may be necessary to re-evaluate and reinterpret the hypotheses presented here based on the findings of my study.

6.2.10. Indirect Evidence for the Breathy/slack/lax Versus Tense/pressed/constricted/stiff Voice Hypothesis

While many of the results for the acoustic analysis support the hypothesis that black speakers make use of a breathier mode of phonation overall, this hypothesis also receives some indirect support from research on isiXhosa. This is based on the findings of Jessen and Roux (2002), under the assumption that a form of transfer effect is at play. As noted in chapter two, Jessen and Roux (2002) were primarily concerned with the use of breathy voice as a means of distinguishing the voiced stops and clicks of isiXhosa.

Jessen and Roux (2002:9-10) observed that for their isiXhosa data, vowels often had their own voice quality contours, such that the first part of the vowel was found to be modal, while the second part of the vowel was to some extent breathy as well as reduced in terms of amplitude. They note that this type of contour was found to dominate all regressive coarticulatory effects of the subsequent consonant (Jessen & Roux 2002:9-10).

Jessen and Roux (2002:26) found that for most of their speakers, the value for H1*−H2* is relatively high around the vowel midpoint and that by this midpoint, the differences in H1*−H2* values linked to different stop and click categories are either eliminated or otherwise reduced. This is again suggestive of a particular vowel specific phonatory pattern for isiXhosa,

independent of consonantal influence. For only one speaker included in their study were there clear differences in terms of H1*−H2* for the midpoints of vowels according to the different stop and click categories. Jessen and Roux (2002:26) note that for isiXhosa, there is an increase in breathiness during the second half of long vowels. These authors link this to a "…voice quality contour often observed in Xhosa" (Jessen and Roux 2002:26).

Observations such as these which point to the use of a language-specific voice quality contour for isiXhosa are potentially of some relevance to the findings of the current study and the hypothesis advanced here regarding voice quality specifically, since all of the black speakers included in my study are of an isiXhosa language background. It is conceivable for example that such a characteristic isiXhosa voice quality contour could be transferred to the English spoken by these participants. Should such a transfer effect be in operation, it would be predicted that higher H1*−H2* values at vowel midpoint for speakers of an isiXhosa language background would be found in comparison to lower values for those who do not have such a linguistic influence, for example, the white speakers included in my study. Likewise, if this trend can be interpreted as an increase in breathiness towards the middle and end points of long vowels for isiXhosa language speakers, a further prediction can be made that values for other measures of breathiness would also be higher. These are precisely the patterns observed in my data and therefore this interpretation would suggest that the differences could be potentially accounted for in terms of a difference in breathiness arising from a type of transfer effect of a commonly used isiXhosa voice quality contour.

However, it is worth noting that Jessen and Roux (2002) did not provide as detailed information about other measurements such as H1*−A3*. For this measure, the pattern at least at vowel midpoint was less consistent than for H1*−H2*. In my study, the H1*−A3* difference is fairly consistent, possibly more so than for H1*−H2* which did not show any clear pattern of difference for the sentence data, whereas this was the case for H1*−A3*.

In addition to the observed voice quality contour for isiXhosa, Jessen and Roux (2002) also observed some voice quality effects linked to the voiced stops and clicks of isiXhosa specifically. They suggest that the main articulatory gesture implicated in the production of the

voiced stops and clicks is larynx lowering which in turn leads to the slackening of the vocal folds with accompanying "glottal leakage" and therefore potentially some degree of concomitant breathy voice (Jessen and Roux 2002:39). This would also lead to a strong lowering of f_0 and some lowering of F1. As already mentioned, the black speakers included in my sample exhibit precisely such lowering of both f_0 and F1 in comparison to white speakers. Potentially therefore there may also be some transfer effect involved in addition to the voice quality contour effect discussed above, namely a larynx lowering gesture in the spoken English of the isiXhosa language background speakers. I propose that the influence of such a transfer effect related to the voice quality linked specifically to the voiced stops and clicks can only partly account for the observed values in my study and perhaps only for some speakers.

Jessen and Roux (2002) describe the voice quality used in signaling the voiced stops and clicks in isiXhosa as 'slack voice.' Ladefoged and Maddieson (1996:63-66) suggest that rather than a configuration involving active arytenoid separation as in breathy voice, slack voice involves loose vibration of the vocal folds without arytenoid separation. This description is partly consistent with some of the patterns observed for black speakers. For example, the values for H1*−H2* would suggest that there may be greater cricoarytenoid tension for white speakers in comparison to black speakers, which would also be the case were the use of slack voice thus described greater among black speakers. The values for this measure would also suggest lesser glottal constriction which would also be a prevailing condition during the production of slack voice. Furthermore, the findings for H1*−A2* and H1*−A3* would suggest non-simultaneous closure of the ligamental glottis for a number of black speakers as well as less abrupt closure which would in fact be consistent with the use of slack voice. However, the fact that black speakers also have higher values for H1*−A1* and that this is one of the more consistent differences and one which shows a clearer division in terms of ethnicity, would suggest that the size of posterior glottal openings among black speakers is an important characteristic of the voice quality produced and this is something which is less likely to be the case without arytenoid separation. I therefore suggest that as far as transfer effects are concerned, the vowel-intrinsic voice quality pattern observed for isiXhosa may potentially be a more important factor than the transfer of 'slack voice' originally linked specifically to the voiced stops and clicks of isiXhosa.

As pointed out by Podesva and Callier (2015:181), research on differences between bilingual speakers has provided compelling evidence regarding language-specificity in terms of voice quality differences, citing Ng, Chen and Chan (2012) who have found that Cantonese-English bilinguals exhibit systematic voice quality differences depending on the language being spoken. Ng, Chen and Chan (2012) for example found that Cantonese exhibits a higher mean spectral energy and lower spectral tilt than English (Ng, Chen and Chan 2012). That such effects have been found for bilinguals provides further support for the idea that a language-specific voice quality contour (potentially originally via some sort of transfer effect) may be able to partly account for the origin of the differences observed in my study between isiXhosa language background bilingual research subjects and the white monolingual English subjects. However, it must be emphasized that this link is speculative, since we have no voice quality data from earlier studies indicating either an increase or decrease in terms of the voice quality parameters included in this study for different groups and no data from longitudinal studies which could verify such an account for SAE.

Giles (1979:260) also cautions against interpreting ethnic differences apparently linked to language differences alone as purely interlingual interference, due to the fact that they can be nevertheless be adopted in a deliberate fashion as ethnic speech markers for the purpose of asserting a distinct linguistic identity. Currently it is unknown whether this applies to this group of speakers or not.

6.2.11. Overall Summary of the Breathy/lax/slack voice Versus Tense/stiff/constricted/pressed Voice Conclusions

To summarize, there is some support for the hypothesis that black speakers make use of a voice quality which overall is characterized by more breathiness or the use of slack or lax voice in comparison to that used by white speakers who may use a generally more constricted, pressed, stiffer and tense voice quality overall based on the acoustic evidence. The exact means whereby this difference is realized cannot be determined with any absolute certainty based on the results for the acoustic measures alone, however these results do suggest a number of mechanisms which may be used with certain gestures being more commonly favoured. This would ultimately require confirmation by means of studies using alternative research techniques some of which are suggested later in this chapter.

My findings are probably most consistent with a scenario in which most of the black speakers have greater posterior glottal openings than white speakers, in addition to nonsimultaneous closure (not necessarily for the same speakers) along the length of the glottis as well as greater open quotient (at least for casual speaking style) and increased glottal stricture on the part of white speakers overall. There is some reason to suggest as discussed above, that nonsimultaneous glottal closure and differences in open quotient are perhaps not as important as differences in posterior glottal opening size, but may nevertheless play a role. Following Zhang's (2016) model, my results support the hypothesis that white speakers employ greater vocal fold thickness than black speakers overall.

Some caution is required in evaluating the hypotheses put forward and discussed in this chapter. There is firstly the matter of classification and whether the observed differences can be characterized as differences which correspond to one of the more traditional voice quality labels. For example, whether the voice quality difference observed is a result of what has traditionally been referred to as breathy voice, or would better be considered 'lax voice' or 'slack voice' (as described by Jessen and Roux 2001 for isiXhosa) is unknown at this point and cannot be known with any certainty based on the acoustic results alone. This is because several of the acoustic characteristics of breathy voice are similar to those of lax voice, the difference being a matter of degree.

As discussed by Garellek (2016), voice qualities are defined in relative rather than absolute terms in comparison to one another, such that for example, certain voices can be described as creakier than others while others are breathier. This is why, according to Garellek (2016), there are many terms which have been used to cover essentially the same range of voice qualities. For example, there is a great deal of overlap between voice qualities characterized by varying degrees of breathiness, including for example, qualities such as 'aspirated,' 'slack,' 'lax' and 'murmured', while voice qualities characterized by creak are described variously as 'tense,' 'stiff,' 'laryngealized,' 'pressed' and 'glottalized' (Garellek 2016).

As pointed out by Garellek (2016), 'lax' or 'slack' voice is occasionally considered to be an intermediate quality between modal voice and breathy voice and 'stiff' or 'tense' voice is likewise considered intermediate between modal voice and creaky voice, citing Keating et al

(2011) and Kuang and Keating (2014). Given the mostly relatively small differences in terms of the trends observed between black and white speakers, I would suggest that it would perhaps be fair to characterize white speakers as having a tenser or stiffer voice quality overall, with a more lax or slack voice quality being adopted by black speakers.

Difficulties in terms of such labels arise primarily due to the difficulty in identifying breathy voice both acoustically and perceptually given that as noted by Gerratt and Kreiman (2001:377), breathiness is a continuous rather than a discrete variable, such that there is a continuum between modal voice and breathy voice which are two types which in practice are difficult to distinguish from one another. As noted by Gerratt and Kreiman (2001:377), many listeners are unable to agree on whether a particular voice is breathy or not, which is reflected in the challenges encountered in the auditory analysis presented in the previous chapter as already described. Thus whether the differences accord well with what is traditionally understood in terms of actual perceptual breathy voice is difficult to determine at this stage, thus inviting perception studies investigating this phenomenon in future.

Perhaps of greater consequence however, Gerratt and Kreiman (2001:378) point out that perceptually, a voice may not sound particularly breathy even if produced with a relatively large glottal gap and voices with relatively larger noise components may also not sound noticeably breathy. The problematic nature of identifying perceptually relevant differences in breathiness is further complicated by the fact that it appears that different linguistic groups may be sensitive to different parameters of perceptual breathy voice (see Esposito 2010 for example).

Thus while the findings of the present study are important in that they indicate that it may be likely that black speakers make use of an habitual voice quality characterized by the greater use of some form of breathy voice in comparison to white speakers who use a tenser setting, ultimately carefully designed perception studies would be needed in order to test the perceptual relevance of the acoustic differences for speakers of South African English. As noted by Garellek (2016:12), citing Garellek, Samlan, Gerratt and Kreiman (2016), listeners are not equally sensitive to all spectral components and in addition, Garellek et al (2013) have found that sometimes the effect of one component of spectral tilt may cancel out the effect of other slope

components perceptually. It is clear from findings by Zhang, Kreiman, Gerrat and Garellek (2013:462) that even in cases where there are fairly dramatic changes, for example, in terms of vibratory regime, it cannot necessarily be assumed that such changes resulting in acoustic events are perceptually important to listeners. Therefore it must be acknowledged that more work needs to be done in terms of perception studies to discover the perceptual relevance of some of these differences. The challenges of such research are discussed in the course of a discussion of the limitations of this study and directions for future research presented in a later section of this chapter.

As stated by Garellek (2016:18), it is not currently clear, based on our present state of knowledge about the perceptual relevance of the acoustic measures, whether representing spectral tilt by means of harmonic based differences (for example, H1−H2, H2−H4, H4−H2 kHz and H2 kHz−H5 kHz as included in the psychoacoustic model proposed by Kreiman et al 2014) or alternatively, by representing spectral tilt by means of formant-based measures (such as H1−A1, H1−A2 and H1−A3) provides a more accurate reflection of how listeners perceive voice quality differences. The trends in ethnic differences for the acoustic measures observed in my study were more robust and distinct for the formant-based measures than for the harmonic-based measures in general, however the results from both types did generally agree with one another in terms of their overall interpretation.

6.3. POTENTIAL SOCIOLINGUISTIC SIGNIFICANCE AND THE ROLE OF VOICE QUALITY VARIATION AS AN INDEX OF ETHNIC IDENTITY

6.3.1. Sociolinguistic Significance

Having discussed the nature of the observed patterns in terms of possible articulatory and aerodynamic correlates in the previous chapter, in this section I discuss various possibilities relating to plausible social meanings attached to the hypothesized differences as well as offering a discussion of their possible status as indices of ethnolinguistic identity.

The social meaning attached to different voice qualities generally varies depending on sociocultural as well as more linguistically motivated factors. For example, as discussed by

Podesva and Callier (2015:176), breathy phonation coupled with low pitch in Lachixío Zapotec is used in creative ways in order to index authority.

Podesva and Callier (2015:176) nevertheless note that there may be cultural constraints which operate on the physical factors in mediating the understanding of the indexical features of the voice. Thus, as mentioned in previous chapters, in a different culture, different phonational contrasts may signal authority, in other words, there is no intrinsic link between the signaling of authority and using low f_0 and breathy voice in every linguistic and cultural group. Thus for example, even though the black speakers in my study display lower fundamental frequency and evidence of breathy voice, this cannot be interpreted to mean that in South African English this necessarily indexes authority in this dialect of English as it does in Lachixío Zapotec. Since the observed differences may have a different meaning in the context of South African English and may even potentially have different meanings for different ethnic groups within South Africa, any conclusions drawn in this regard must necessarily remain speculative for the time being.

In terms of my own research findings specifically as already alluded to, even if perceptual studies and research using other methodologies and techniques establish that the acoustic results can be linked to the greater use of perceptual breathy voice by black speakers, it would nevertheless remain unclear as to whether this would represent an increase in the use of breathy voice over time or a decrease. This is because there are no prior studies of similar groups of speakers in South Africa investigating the use of different voice qualities which could serve as a basis for comparison. Thus for example, it is just as likely to be the case that white speakers have adopted a voice quality more characterized by greater glottal stricture as well as thickness and stiffness, whereas in the past they may have used breathy voice habitually. Clearly more work along these lines is needed as will be suggested in a later section of this chapter.

Szakay (2008:43) cites Gussenhoven (2002) as claiming that higher pitch is frequently associated with politeness, vulnerability, femininity and submissiveness, whereas there is an association between lowered pitch and masculinity, assertiveness and authority. Henton and Bladon (1985) note the association between breathiness and ‘attractiveness’ or ‘arousal,’ suggesting that the habitual use of breathy voice by female speakers of certain British English dialects may be used to increase approbation by male interlocutors. Sicoli (2007), cited by Podesva and Callier (2015), found that authority was creatively indexed by Lachixío Zapotec

speakers by using breathy phonation and lower pitch. As noted by Podesva and Callier (2015) however, cultural constraints mediate the process whereby voice characteristics come to be indexical of "physical realities." Thus the question of what the exact indexical meaning of the lower fundamental frequency observed for black speakers in my study is for speakers of South African English is as yet unknown and invites further research. Firstly it is possible, given the discussion provided above that the differences in f_0 may simply be a concomitant effect of black speakers using breathy voice more habitually. Secondly, it is not clear whether the use of low or high pitch level carry the same meanings for black and white speakers. Thirdly, this may be as a result of some kind of language-specific transfer effect without necessarily having acquired a particular social meaning at present.

6.3.2. Voice Quality and Ethnicity

In this section, I explore the potential relation of the hypothesized differences as they relate to questions of ethnic identity. As pointed out by Podesva and Callier (2015:180), most work which has investigated the connection between voice quality and ethnicity has focused on African American speech. The research that has been conducted to date has suggested that, according to Podesva and Callier (2015:180) citing Irwin (1977) and Thomas and Reaser (2004), voice quality cues may play an important role in speaker race identification at least for American English speakers. Research by Alim (2004), Britt (2011) and Moisik (2013) has shown that a variety of non-modal phonation types, such as a type of falsetto as well as harsh voice are used and are associated with the African American community (Podesva and Callier 2015:180). In my study according to the auditory analysis of the sentence data, black speakers were found to use non-modal phonation types at greater frequencies than their white counterparts, although the difference according to this analysis was marginal.

However, when considering whether the observed differences in terms of voice quality could be indexical of ethnicity, it is also important to consider another alternative, namely that they may be indexical of a linguistic group rather than an ethnic group. Podesva and Callier (2015:180) for example, point out that linguistic identity can be indexed by voice quality and note that it is probable that both languages as well as dialects may differ from one another in

terms of voice quality[57], possibly due to the influence of sociocultural factors. As Ng, Chen and Chan (2012:e171) point out, because it is often the case that speakers of differing ethnicities also happen to speak different languages to one another (and this was certainly the case for my sample where none of the white speakers included in the sample were in any way proficient speakers of isiXhosa), vocal differences could conceivably be linked to language rather than ethnicity in such cases. Because the black speakers in my study were all of a sub-ethnic Xhosa group, it would be necessary to sample black speakers of different sub-ethnic groups who speak the same language in order to assess the relative contributions of language background and ethnicity respectively. This would necessitate the investigation of the effect of language on differences in for example, fundamental frequency (Ng, Chen and Chan 2012:e171).

As noted by Ng, Chen and Chang (2012:e171) with particular reference to their discussion of differences in fundamental frequency between speakers (although the discussion is relevant to wider differences in voice quality generally), if a difference in mean fundamental frequency is observed between two groups of speakers who differ both in terms of linguistic background as well as ethnicity, the observed difference can be attributed either to some underlying anatomical difference between the two groups or due to the influence of the different languages in terms of differences in average fundamental frequency. Thus for example, in my study where there are differences between the two groups in terms of f_0 (as well as for other measures), it is possible that these differences can be attributed to potentially either anatomical differences or to linguistic differences and consequently, were these differences to index a particular identity, this could potentially be either linguistic identity (as already noted, potentially arising from differences in terms of the use of English as part of a repertoire including an African language versus English being used as a main L1), in which case, the difference would be likely to be motivated by language transfer[58] or social convention alone, but if ethnic differences, these could be both physically motivated or established by social convention.

As Mennen, Scobbie, de Leeuw, Schaeffler and Schaeffler (2010:26) note, pitch range settings can be used as sociolinguistic markers signaling typical vocal tendencies associated with

[57] The South African context entails further complex considerations. The observed patterns might not be amenable to descriptions purely in terms of a difference between dialects, but rather in terms of the difference between having English as a main L1 as opposed to English being included as part of a multilingual repertoire.
[58] The degree of influence of any African language substrate may vary according to other factors such as social class as well as the status of English as either an L1 or an L2 for a given speaker.

accents and languages. In addition, Mennen et al (2010:26) mention that numerous studies have revealed consistent differences in pitch level between different languages even after factoring out individual differences, citing Braun (1995). The differences in pitch span observed in my findings may therefore be indicative of language-specific pitch differences that have been carried over from isiXhosa, although more research would be needed to evaluate this possibility.

Laver (1968:44) states that there are two principal sources which voice quality derives from, namely the anatomical source and secondly the laryngeal as well as supralaryngeal 'settings' which can be acquired either by means of social imitation or idiosyncratically and which the speaker may be unconscious of once acquired. The range in which the long-term settings operate, according to Laver (1968:44), is constrained by the speaker's basic physiology and anatomy. Taking this into consideration therefore, Laver's (1968) discussion allows for any anatomical differences between speakers which may exist to operate in a way which imposes a type of upper limit on the range in which long-term settings can operate. While not explicitly mentioning ethnic differences in terms of vocal tract anatomy, Laver's (1968) discussion thus does allow for this possibility. In his discussion, Laver (1968:48) groups the indexical possibilities of voice quality under three main categories, namely biological, psychological as well as those providing social information. While sociolinguists are primarily concerned with the social indexical function of voice quality, the biological indexical function, as described by Laver (1968:48) is also important to take into account even though it derives from factors outside of the control of the speaker, namely physiology.

This point is particularly pertinent given that there is little research, as mentioned in chapters one and two, which clearly demonstrates a definite link between differences in the acoustic measures included in this study and specific anatomical differences for the groups in question. Thus while the evidence is still lacking to establish beyond any doubt that the observed patterns could be attributed either to anatomical differences or to a particular voice quality indexing a certain ethnic or linguistic identity, such differences are at least potentially available to function in such a way.

Using Laver's (1968) terminology, if the source of the voice quality differences are anatomically-derived, they can be regarded as 'intrinsic' voice quality features since they are in

that case not under the voluntary control of the speakers in question. This being opposed to 'extrinsic' voice quality features which the speaker is able to control at will. Thus if it can be established that the differences in terms of the tendencies I have observed in my study have a purely organic, anatomical base, then we could say that their effects on voice quality can be considered intrinsic voice quality features. Alternatively, if no purely anatomical difference can be confirmed they could be considered as extrinsic voice quality features.

In South Africa, it is clear that there are very few studies which deal with such questions relating to interethnic anatomical differences in vocal tract structure and size directly and therefore at this stage, anatomical differences cannot be entirely ruled out as one of the possible contributors towards the observed probabilistic tendencies with regards to the acoustic measures found I my study[59].

Wells (1982:92) at least implicitly allows for the possibility that ethnic differences in speech may have an anatomical source, in cases where the speakers involved are from genetically distinct populations, as is the case for my participants. Laver (1968:51) notes that it would be of value to have anthropometric measurements which could provide useful information about typical variations in anatomical dimensions in addition to the acoustic data such as that presented in this study. The result is that comments regarding whether differences such as those observed in my study can be considered to be the result of intrinsic or alternatively extrinsic differences in voice quality must remain speculative for now.

Even if differences in anatomy can account for many of the observed patterns, it is clear that these need not be discrete, as there are a number of exceptions to the general trends as discussed below, possibly reflecting a range of responses along a continuum. There is substantial interspeaker variation in terms of the use of the different parameters even within ethnic groups.

[59] One of the few studies directly investigating such interethnic differences is Boshoff (1945), a study which compared the morphological structure of excised cadaveric larynges from black and white South Africans. While Walton and Orlikoff (1994) cite this study in their discussion of ethnic differences in American English, they note that social factors are more important in terms of ethnic dialect variation than those linked purely to anatomy. Kreiman and Sidtis (2011) reach a similar conclusion in terms of the relative importance of dialectal as opposed to anatomical differences. This issue is discussed in greater detail in chapter two of this thesis.

Thus there are certain white speakers who exhibit higher values for some of the measures for which most black speakers overall are found to have higher values and vice versa. For example, while black speakers overall display higher values for H1*−A1*, the second lowest mean values for this measure were found for a black speaker. Likewise, the third highest average value for this measure was that of a white speaker. It is therefore clear that while the observed differences form trends in terms of ethnicity, there is firstly a fair degree of variability in which these trends are expressed and furthermore, that there are individuals who do not closely follow the trends for their respective ethnic groups. In terms of social characteristics, these speakers do not differ greatly from most other members of their respective ethnic groups as far as I have yet been able to discern based on information provided during the interviews. However there are certain patterns which display the individual nature of differences in voice quality which can be accounted for in terms of individual vocal profiles.

For example, speaker C5d, a white speaker, has relatively low values for CPP (an average of 17.88 dB), whereas the overall tendency for white speakers is to have higher values for this measure. However, this speaker also shows some of the lowest average values for H1*−A3* (14.280) and H1*−H2* (-0.5146 dB) and H1*−A1* (19.100 dB), in line with the general trend for her ethnic group for these measures. This would suggest that this particular speaker has a high degree of aperiodicity, but low open quotient (or higher glottal stricture), a smaller posterior glottal opening and more abrupt closure of the vocal folds. When listening to this speaker, it is clear that she consistently uses creak throughout most of her utterances to a degree perhaps greater than most white speakers. Thus the lower CPP values in the case of this speaker can presumably be accounted for not in terms of breathiness as for most black speakers, but in terms of the modulation noise introduced by creaky phonation. There are also exceptions to the general trends amongst black speakers. Speaker M1 for example, exhibits the second lowest H1*−A1* values (19.510 dB) and the fourth lowest values for H1*−A2* (22.34 dB) contrary to the trend towards higher values for these measures amongst black speakers. The auditory impression in listening to this speaker is that of someone who uses a fair degree of modal voice and constricted forms of creak to some extent and does not sound particularly breathy. There are also some white speakers who exhibit higher values for some measures for which the trend is for black speakers as a group to have higher values. For example, J3, a white speaker, has the second highest values

for H1*−A1* (27.48 dB), H1*−A2* (30.02 dB) and for H1*−A3* (25.720 dB). The auditory impression of this speaker is one of a person whose voice has a softer and breathier quality than most, while at the same time making use of creak utterance finally. That there are exceptions to the general observed trends provides support for the notion that voice quality for this sample of South African English speakers at least, does not function in any kind of deterministic way, but rather in terms of probabilistic trends.

Giles (1979:266) explains that the origins of "intralingual speech markers" are often the result of either involuntary or alternatively voluntary substrate influences as well as possible ensuing hypercorrection. Giles (1979:266) notes that most of these markers are probabilistic as well as continuous in terms of form. By probabilistic, Giles (1979:266) means that they do not occur for every member of the ethnic group under consideration and also may not be used on every occasion by an individual speaker within that group. While my research findings point to a fair degree of consistency in terms of the acoustic patterns across both contextual styles as well as overall regardless of sentence position and prominence, it is clear that not all speakers of either ethnic group adhere to the observed overall norms for their groups and thus it is best to describe these trends as essentially probabilistic in the sense that Giles (1979) uses the term. As pointed out by Giles (1979:281) citing Ryan and Carranza (1977), because ethnic groups do not constitute "homogenous wholes," it is quite possible that some individuals will not follow the norms typical of their respective ethnic groups.

It is also important to provide some assessment of whether the observed differences in my study could potentially perform or are in fact already performing a socially indexical function. Laver (1968:50) states that voice quality is capable of indexing group membership, either of an accent group or in fact, citing Fay and Middleton (1939), groups which are not necessarily defined by accent. This pertains to the features of voice quality, which according to Laver (1968:50), can be imitatively acquired, such that voice quality can function as a clue to typical social characteristics which speakers of that particular accent may exhibit.

Laver (1975:313-314) points out that due to the fact that for the most part, social behaviour is learned, indexical voice quality features to the extent that they are indexical of social information are those which have the capacity to be acquired by means of imitation. This

applies for the most part to extrinsic features (Laver 1975:313-314). Laver (1975:313-314) notes that those extrinsic settings forming characteristic features of a given accent are then able to act as clues to the particular social factors which one typically finds for speakers of the accent in question.

Another way of stating this is to say that in order for voice quality settings to function as indices, there are at least two conditions which must be satisfied, specifically they need to have a capacity to be acquired by means of imitation. Secondly, they need to be consistent enough in order to form characteristic features of a given accent and finally, having become part of the collection of characteristics common to an accent, they gain the propensity to act as clues to social factors, that is, to act as indices of a group of speakers.

The first question to ask then is whether the hypothesized differences between black and white speakers can be considered extrinsic, that is whether they are able to be acquired by means of imitation. The answer to this question must certainly be affirmative as evidenced by the fact that several speakers who are ethnically white are able to achieve breathy modes of phonation and several speakers who are black typically use more pressed/stiff modes of phonation habitually, at least as suggested by the acoustic measures and auditory impression of these speakers overall. There is certainly no evidence of any underlying anatomical impediment to speakers from either ethnic group preventing them from performing the necessary muscular adjustments (compensatory, if need be) in order to effect the hypothesized differences in voice quality.

The second question, namely, whether such differences are consistent enough to form part of an accent which can be recognized as such is at this point more difficult to answer. The evidence provided in this study suggests that the overall trends are quite consistent, but that there is substantial variation between speakers. This in turn would suggest that such characteristics do at least have the potential to perform an indexical function with regards to indexing ethnicity or linguistic group membership (subsumed by the term “ethnolinguistic” in this study). The question of whether the hypothesized differences are both perceivable and recognized by South African listeners is ultimately a question which will need to be answered using appropriate perception studies based on these findings.

6.4. RESEARCH IMPLICATIONS

In this section, I discuss the wider implications and impact of my research in the field of voice quality studies in general and for research on South African English.

One of the wider implications of my research relates to the understanding of variation in voice quality generally and how this relates to variation on an individual level. As stated by Henton and Bladon (1988:23), before one can confidently allocate paralinguistic meanings to phonetic features, it is important to take full account of how such events function as part of an accent. For accents of South African English this is particularly important, since there is little work currently which focuses on the way in which voice quality differences may function for these accents. My research has made a start in this direction.

The findings of this research can potentially provide useful data for the field of speech therapy in South Africa, where norms for different ethnic and linguistic groups have not been convincingly established and while this has not been the primary aim of my research, the findings may nevertheless be helpful in terms of exemplifying the need to consider variation in voice quality as an essential component of accent and therefore in evaluations used in speech therapy work.

Likewise, the current research as well as research emanating from my findings may subsequently prove helpful in the field of forensic linguistics which is currently without sufficient detailed and recent phonetic accounts of voice quality variation in South African English and may thus lead to more accuracy in speaker identification in future.

One of the potential implications of my findings which I have not discussed thus far is how they relate to ongoing studies of vowel quality for accents of South African English as well as other studies which rely heavily on formant and fundamental frequency measurements. My study was not designed to deal with these questions directly, but some of the findings may be of relevance to such studies. In particular, the results of the sentence data, where vowel quality was controlled for, suggested that, in agreement with studies by Iseli, Shue and Alwan (2006, 2007), a number of the measures are affected by vowel quality. In general, for these measures the effect of vowel quality was primary, while the ethnic differences were clearly observable within vowel

categories. This would potentially suggest at least that for the most part, differences in vowel quality are greater than differences in voice quality for this sample overall.

Perhaps of greater interest is the fact that a consistent difference was found between black and white speakers in terms of fundamental frequency. While some studies, mostly investigating intonational phenomena in South African English have documented differences in f_0 for certain groups of speakers linked to prominence effects, my research findings suggest that there are perhaps more global differences in fundamental frequency for such groups than previously anticipated. The fact that differences in fundamental frequency were observed, that is, across both non-prominent as well as prominent syllables and in different contextual styles may simply be due to the fact that the effect of f_0 lowering linked to prominence for black speakers is so great as to have an overall effect when measuring both prominent and non-prominent syllables together. However, it could also mean that, particularly if the breathy/slack versus stiff/tense hypothesis set forth in this thesis is correct (and the f_0 lowering observed is a function of this difference) that the f_0 effect is a result of voice quality differences rather than prominence effects. The question of which factors may be most important in contributing to the observed f_0 difference may well be worth exploring in future research.

Ultimately, my research findings may also contribute to the development of more realistic speech synthesis (including its application in the field of speech technology) for South African English speakers by way of the fine phonetic detail provided here.

6.5. LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

In this section, I discuss the limitations of the findings of this study, and in so doing provide suggestions for further research questions and hypotheses which have yet to be tested which will provide directions for future research in this area. Due to the exploratory nature of my research, many questions remain unanswered and I will also address such issues in this section.

I first discuss those limitations relating to the research design and methodology. One of these limitations is linked to the specific judgement sample used and therefore the characteristics

particular to this sample of speakers. For example, all of the black speakers included in my sample are of an isiXhosa language background. As explained in chapter three, this was necessary to control for the possibility that, should there be any differences in phonation present between black and white speakers stemming from language-specific transfer effects (a distinct possibility given the foregoing discussion), by not controlling for language background, any such effects would be potentially obscured. Thus it is not possible to extend the findings of my study to all black speakers of South African English. Likewise, it is not possible to generalize the research findings to apply to speakers of South African English who are not English-dominant. Furthermore, it may be necessary to do more work of this nature focused on languages in South Africa other than English, such as isiXhosa, in order to shed light on the role of language-specific phonatory differences and potential transfer effects, of the type suggested by Ng, Chen and Chang (2012:e175).

Another limitation arising from the sample selection procedure is that because all of the research subjects are female, the research findings cannot be generalized to the male population of either ethnicity. Firstly, as noted in chapter two, a number of researchers have reported differences in phonation between males and females thought to be linked to anatomical differences between males and females (Gordon and Ladefoged 2001, Esling and Edmonson 2011). In addition, it would not be possible to compare male and female speakers in terms of differences in breathiness for example, using certain measures such as H1−H2, as explained by Simpson (2009). Nevertheless, investigating differences in voice quality linked to gender in the South African context is a potentially fruitful avenue for future research although any study of this nature including both male and female speakers would need to include careful attention to suitable research design and appropriate measures to be used.

Towards the beginning of this thesis, I posed the question whether any differences in terms of the acoustic measures that might be found between the black and white speakers included in this study could be the result of certain differences in voice quality having been assigned an important role in ethnic identity formation and hence ethnic identification among South African English speakers. However, as Podesva and Callier (2015:174) insightfully note, following Bucholtz and Hall (2005), social categories such as race, ethnicity and gender are only

one of many possible relevant positions a subject may adopt in terms of identity construction (Podesva and Callier 2015:174). That is, it is important when evaluating the role of voice quality differences in identity construction that the relevance of gender as well as other social categories should also be explored in order to test the importance of these other categories. The current study has provided a necessary baseline in order to further evaluate the contributions of other social categories to the manifestation of voice quality differences linked to identity in South African English in future research.

My study has relied mostly on production data rather than perception data as such and cannot therefore offer direct insights into the perceptual relevance of the patterns observed. Podesva and Callier (2015:180) point out that the extent to which different communities are sensitive to the distinctions in terms of voice quality based on race may often vary. They note for example that while such distinctions are relatively salient to American English listeners, research in other English-speaking communities reveals differences in the extent to which listeners are accurate in identifying speaker ethnicity. In this regard, Podesva and Callier (2015:180) cite Tan (2012), who found that older Singaporean listeners are able to identify speaker ethnicity more accurately and consistently than younger Singaporeans which Tan (2012) attributes to several factors, including exposure to and experience of different accents, as well as government policy effects resulting in a weakening of ethnic consciousness among the youth of Singapore.

Given therefore that it is also possible that even within the same language or dialect different ethnic groups for example may attach different meanings to the observed differences, a potentially worthwhile topic for further investigation would be whether these same differences observed in terms of the use of hypothesized breathiness and low f_0 may index different meanings for different groups within South Africa. That different ethnic groups may make use of different cues or attach more importance to certain cues in comparison to others in speaker ethnicity identification tasks is perhaps suggested by the work of Mesthrie, Chevalier and McLachlan (2015). This research revealed that not all ethnic and gender groupings were equally successful in distinguishing between ethnic groups based on accent features (although it should be noted that the importance of differences in voice quality relative to vowel quality in perceptual terms for this group of speakers is not yet known) possibly due to differences in terms of exposure and experience as suggested above.

In the South African context, isolating the effect of language poses a particular challenge, since it may be difficult to locate bilingual white speakers who are equally proficient in isiXhosa as black speakers who are proficient in English, following the methodology suggested by Ng, Chen and Chang (2012:e172) to address such a question. However, if the interest is particularly focused on the meaning which is assigned to such differences, perceptual studies as well as the use of synthesized speech materials may be a more suitable option. Studies of bilinguals where dominance of either English or isiXhosa could be carefully controlled for would be one possible solution for examining the contribution of language to the differences observed in this study, as Ng, Chen and Chang (2012:e174) for example suggest that speaking fundamental frequency in one language may be affected by the competency in another language at least for female speakers.

As the methods used here are primarily acoustic, not much can be directly inferred regarding articulation. Acoustic measures were chosen as the focus for this study, for several reasons, including accessibility, cost, the fact that methodologies for acoustic measures are reasonably well-established, their non-invasive nature and the facility with which such data can be computed (Mennen et al 2010:26). Nevertheless, there are inherent limitations to the use of such methods and these therefore also apply in the case of my study. One of the limitations of using such methods is that acoustics do not always map in a corresponding one-to-one fashion with articulation as well as in terms of speech perception (Mennen et al 2010:26). Thus the articulatory possibilities discussed in this chapter as well as the perceptual correlates can only be inferred from prior research. Secondly, while it is hoped that because the measures used in this study included corrections for formants and bandwidths and thus most of the effects of supralaryngeal articulation are mostly controlled for, it is possible that some of the acoustic measures may be influenced by factors relating to the filter in addition to the source. Fortunately, as pointed out by Mennen et al (2010:26), valid articulatory models and perceptual models can reduce the degrees of freedom necessary when using acoustic methods and can aid in forming more valid interpretations. Appropriate measures are in constant need of improvement (Mennen et al 2010:35) and given a greater research interest in voice quality internationally, continuing

work persists in contributing towards useful models which can be applied to research findings such as those of this study.

As Garellek (2016:10) points out, because not all acoustic parameters included in current voice articulation models may be relevant for perception and it is possible that such models are not capable of accounting for every possible perceptual voice quality change, both acoustic characteristics as well as articulatory correlates are only relevant to voice quality to the extent that they play a role in the perception of different voice qualities. As noted earlier, scholarship in this area would benefit greatly from an improved understanding of the perceptual relevance of the various acoustic measures included in this study for South African listeners (for the measure H4−H2K in particular). Perception studies focusing on these issues specifically would therefore be particularly helpful.

Kreiman et al's (2014) psychoacoustic model as well as modeling studies such as those of Zhang (2016a) contribute to a greater understanding of how acoustic measures relate to underlying vocal fold biomechanics and physiology as well as perceptual differences in voice quality. As these models are continually being iteratively refined by means of systematic investigations of the various parameters involved, it will be possible to establish "a cause-effect theory of voice production" (Zhang 2016b:2631). Such theories will enable future sociophonetic work to make increasingly accurate and precise interpretations of acoustic data relating to voice quality variation in future.

As mentioned by Esling (2006), cineradiographic, electromyographic, cinephotographic, endoscopic, electropalatographic, fibre optic as well as aerodynamic techniques and high-speed laryngoscopy may be used to supplement findings such as those made in this study.

One of the limitations of this study is that, given that there have not been prior studies demonstrating the use of different voice qualities in South African English before, it is impossible to say what the role of voice quality is in linguistic change at this point in time. Thus future research could focus on providing either longitudinal studies or apparent-time studies using different linguistic and ethnic groups to shed light on the question of whether differences such as those which I have discovered are increasing or alternatively decreasing. Such research may therefore potentially also provide us with useful clues as to the meaning of such differences in voice quality in South African English and their sociolinguistic function.

6.5. CONCLUSION

In conclusion, I have isolated, described and analyzed ethnolinguistic differences in voice quality using acoustic techniques and a supporting auditory analysis. The findings of this analysis suggest certain specific hypothesized differences between white and black speakers (specifically those of an IsiXhosa language background), namely that white speakers make use of a voice quality characterized by greater vocal fold constriction, thickness and stiffness overall when compared to black speakers who are hypothesized to use a voice quality which may be described as involving greater breathiness. In as far as the effects of vowel quality could be controlled for, evaluated and tested, the results of this study indicate that the observed differences in voice quality may indeed be present even for speakers who are similar in terms of segmental pronunciation. I have also offered suggestions for future research based on these results. I have, by advancing our knowledge of ethnolinguistic voice quality variation, contributed towards a more complete understanding of sociolinguistic variation in South African English.

APPENDIX A: List of Sentences

Say uh, pea again.

Say uh, tea again.

Say uh, key again.

Say uh, bee again.

Say uh, Dee again.

Say uh, ghee again.

Say uh, he again.

Say uh, fee again.

Say uh, she again.

Say uh, see again.

Say uh, thee again.

Say uh, Vee again.

Say uh, Zee again.

Say uh, chee again.

Say uh, gee again.

Say uh, ye again.

Say uh, we again.

Say uh, re again.

Say uh, Lee again.

Say uh, me again.

Say uh, knee again.

Say uh, E again.

Say uh, peat again.

Say uh, teat again.

Say uh, keet again.

Say uh, beat again.

Say uh, Deet again.
Say uh, heat again.
Say uh, feet again.
Say uh, sheet again.
Say uh, seat again.
Say uh, thete again.
Say uh, Veet again.
Say uh, cheat again.
Say uh, geat again.
Say uh, yeat again.
Say uh, wheat again.
Say uh, reet again.
Say uh, leat again.
Say uh, meet again.
Say uh, neat again.
Say uh, eat again.

Say uh, port again.
Say uh, caught again.
Say uh, taught again.
Say uh, bought again.
Say uh, daut again.
Say uh, ghaut again.
Say uh, haught again.
Say uh, fought again.
Say uh, short again.
Say uh, sought again.
Say uh, thought again.
Say uh, vaut again.
Say uh, wort again.
Say uh, wrought again.

Say uh, mort again.
Say uh, naught again.
Say uh, ought again.

Say uh, paw again.
Say uh, caw again.
Say uh, taw again.
Say uh, bore again.
Say uh, daw again.
Say uh, gaw again.
Say uh, haw again.
Say uh, fawe again.
Say uh, shaw again.
Say uh, saw again.
Say uh, thaw again.
Say uh, chaw again.
Say uh, jaw again.
Say uh, yaw again.
Say uh, war again.
Say uh, raw again.
Say uh, law again.
Say uh, maw again.
Say uh, gnaw again.
Say uh, awe again.

Say uh, pun again.
Say uh, cun again.
Say uh, ton again.
Say uh, bun again.
Say uh, done again.
Say uh, gun again.

Say uh, Hun again.
Say uh, fun again.
Say uh, shun again.
Say uh, son again.
Say uh, jun again.
Say uh, yon again.
Say uh, one again.
Say uh, run again.
Say uh, mun again.
Say uh, nun again.

Say uh, cut again.
Say uh, putt again.
Say uh, tut again.
Say uh, but again.
Say uh, dut again.
Say uh, gut again.
Say uh, hut again.
Say uh, shut again.
Say uh, zut again.
Say uh, jut again.
Say uh, yut again.
Say uh, rut again.
Say uh, mutt again.
Say uh, nut again.
Say uh, ut again.

Say uh, foot again.
Say uh, hit again.
Say uh, het again.
Say uh, hat again.

Say uh, hot again.

Say uh, heart again.

Say uh, hurt again.

Say uh, hate again.

Say uh, hoe again.

Say uh, hoot again.

Say uh, height again.

Say uh, Hoyt again.

Say uh, out again.

Say uh, hear again.

Say uh, hair again.

Say uh, who’re again.

Say uh, hutter again.

APPENDIX B: Routinely Elicited Interview Questions

1. Do you consent to taking part and being recorded (following verbal explanation of procedures and ethical considerations)?
2. What is your name?
3. What is your date of birth?
4. Where were you born and where did you grow up?
5. Where do you live now (which suburb)?
6. How long have you been living there for?
7. How would you describe the area?
8. Have you ever been overseas and if so, for how long?
9. Where else have you lived and for how long?
10. Do you have any brothers or sisters?
11. Where do they live/work/go to school?
12. Do any other relatives of yours live in Cape Town?
13. Where were your parents born and raised?
14. Do you speak any other languages aside from English at home?
15. Have you always used mostly English at home?
16. If not, do you recall when you started using mostly English?
17. Which language do you feel most comfortable conversing in (if any other languages mentioned aside from English)?
18. Do you speak any other language or languages with brothers and sisters, other relatives, grandparents or friends?
19. Do your parents speak any other language or languages?
20. What work do your parents do?
21. Which primary school did you attend?
22. Did you have lots of friends in primary school? Are you still friends with any of them now?
23. What is your best memory of primary school?
24. Which high school did you attend?
25. What was the best thing about high school?
26. Do any of your friends live in the same neighbourhood as you?
27. Did you develop friendships outside of school?
28. Would you say that the high school(s) which you attended was/were well racially integrated?
29. Would you say that the primary school(s) you attended was/were as racially integrated?
30. Did you ever speak any language(s) at school aside from English?
31. Which subjects did you study at school?
32. What were your favourite school subjects?
33. Which subjects didn't you enjoy?

34. Do you think things have changed in schools since the days your parents went to school and if so in what ways?
35. Have you ever been a victim of crime?

APPENDIX C: Descriptive Statistics for the Interview Data

The following table displays the means ($\bar{x}$), standard deviations (σ), medians, maxima, minima, first (1st Qu.) and third (3rd Qu.) quartiles, total number of tokens for each sample (N) and the number of tokens for which VS did not provide a measurement (NA's) for each of the VS measures (corrected and uncorrected) for each of the two ethnic groups.

Measure	**Min**	**1st Qu.**	**Median**	$\bar{x}$	**3rd Qu.**	**Max**	**NA's**	σ	**ethnicity**	N
H1*−A1*	-5.262	18.29	24.337	23.992	30.133	47.153	632	8.432098	black	4066
	-4.928	17.256	23.189	22.608	28.435	48.464	748	8.534643	white	5266
2K*−5K	-4.772	17.09	22.678	22.542	27.511	47.577	632	8.38247	black	4066
	-7.632	15.71	21.392	21.253	26.767	47.453	748	8.594691	white	5266
H1*−H2*	-20.444	-1.272	7.577	4.232	11.128	26.817	608	10.23337	black	4066
	-20.361	-1.1	6.312	4.194	10.619	26.707	733	9.607062	white	5266
H4*−2K*	-19.38	1.7	7.184	7.267	13.543	31.881	632	9.442619	black	4066
	-19.393	2.274	7.168	7.255	12.576	31.89	748	8.971258	white	5266
H2*−H4*	-14.04	1.268	5.848	6.446	11.708	24.934	608	7.429309	black	4066
	-14.179	0.6575	4.876	5.6088	10.766	24.917	735	7.561845	white	5266
H1*−A2*	-0.411	22.298	28.302	27.572	32.962	52.456	632	7.788957	black	4066
	-6.765	20.942	26.78	26.108	31.481	52.575	748	7.988167	white	5266
H1*−A3*	-10.45	14.6	21.21	20.83	26.96	44.37	632	9.290312	black	4066
	-10.6	13.8	19.4	18.93	24.46	44.17	748	8.30816	white	5266
CPP	12.85	18.7	21.46	21.48	24.19	33.92	774	3.609663	black	4066
	12.69	20.27	23.8	23.21	26.49	33.58	895	4.304102	white	5266
HNR05	1.8	23.55	34.4	33.31	43.23	66.74	774	12.60292	black	4066
	-0.571	28.888	42.699	40.255	51.472	71.006	895	14.01193	white	5266
HNR15	6.759	33.174	42.193	41.319	49.981	75.973	774	11.60987	black	4066
	11.59	37.47	48.91	47	57.23	80.21	895	13.1291	white	5266
HNR25	14.13	38.85	46.26	45.72	53	81.08	774	10.38153	black	4066
	17.56	42.75	51.94	51.14	60.15	84.12	895	11.81378	white	5266
HNR35	17.54	41.45	47.76	47.27	53.59	76.92	774	8.856524	black	4066
	22.88	45.28	52.98	52.45	60.12	80.79	895	10.30785	white	5266
SHR	0.001	0.043	0.1	0.1501	0.2	0.932	2248	0.153839	black	4066
	0.001	0.052	0.149	0.2162	0.312	0.979	2545	0.211587	white	5266
pF0	42.2	107.2	163.3	153.7	188.2	302.9		49.39526	black	4066
	42.05	122.78	184.56	170.03	211.79	301.65		56.08028	white	5266
pF1	162.4	393.3	481.1	501.2	599.1	961.1		144.5741	black	4066
	210	393.8	521.5	534	657.6	973.9		171.0046	white	5266
pF2	598.6	1564.7	1945.6	1962.8	2380.1	3286		526.6733	black	4066
	627.2	1567.9	1894.3	1943.4	2362.8	3325.9		514.6268	white	5266

pF3	1934	3119	3395	3415	3725	5408		469.5971	black	4066
	2143	3199	3466	3487	3771	5363		434.0352	white	5266
pF4	3259	4426	4634	4678	4901	5931	328	366.7526	black	4066
	3289	4443	4663	4712	4930	5944	602	393.9294	white	5266
H1−A1	-15.477	-0.502	3.99	3.614	8.682	19.598	633	6.867423	black	4066
	-15.488	-1.866	2.214	2.289	7.154	19.39	749	6.746666	white	5266
2K−5K	-0.916	7.286	8.876	10.179	12.516	21.131	1717	3.936922	black	4066
	-5.303	7.159	8.508	9.44	11.489	21.072	2165	3.746077	white	5266
H1−H2	-20.773	-3.0625	3.036	0.1331	5.6055	19.197	603	8.507062	black	4066
	-20.773	-3.101	2.639	1.053	6.644	20.629	732	8.810641	white	5266
H4−2K	-2.11	15.35	21.48	22.08	28.24	46.9	603	8.883487	black	4066
	-3.845	14.258	21.673	21.434	28.431	47.027	731	9.049653	white	5266
H2−H4	-12.366	4.264	8.226	9.284	14.105	26.888	603	7.365135	black	4066
	-13.135	1.494	7.194	7.856	14.056	26.586	732	8.279897	white	5266
H1−A2	-10.03	16.37	22.55	21.98	28.22	41.48	632	8.697804	black	4066
	-10.19	13.68	20.01	19.81	26.46	40.66	749	8.933336	white	5266
H1−A3	4.399	27.219	32.815	31.725	37.221	45.641	632	7.15555	black	4066
	3.942	26.968	31.907	31.186	36.438	46.716	748	6.690168	white	5266

APPENDIX D: Dialect History

The following table displays the dialect history of the speakers included in this study (in terms of city of origin).

Speaker	Race	Dialect History
A3d	white	Johannesburg
C5d	white	Pietermaritzburg
J1	white	Cape Town
J2	white	Johannesburg
K2	white	Durban (0-10); Pietermaritzburg (10-)
L2	white	Pietermaritzburg
M2	white	Cape Town
M3	white	Cape Town
R3	white	Cape Town
E3d	white	Namibia (0-2); Cape Town (2-)
D1d	white	Cape Town
TOG	white	Johannesburg
E1	white	Cape Town
C6d	white	Cape Town
C1	white	Cape Town
J4d	white	Johannesburg
E2	white	Cape Town
J3	white	Johannesburg (0-13); Durban (13-)
S2	black	East London
S3	black	Cape Town
T1	black	Umtata (pre-primary); Cape Town (since primary school)
T2b	black	New York (0-6); Johannesburg (6-)
B2	black	Port Elizabeth
Z1	black	Queenstown
K	black	Johannesburg
P1	black	Port Elizabeth
S1	black	Johannesburg (0-12); Cape Town (12-13); Hobhouse (13-18); Cape Town (18-)
N2	black	Queenstown
N3	black	Cape Town
M1	black	Port Elizabeth
T3	black	Cape Town
A2	black	Johannesburg (till end of primary school);Bloemfontein (since primary school)
A1	black	Cape Town (0-4); Johannesburg (4-)
O2	black	Kingwilliamstown
B1	black	Margate
A4	black	Bizana (boarding school in Cape Town)

APPENDIX E: Scatterplots

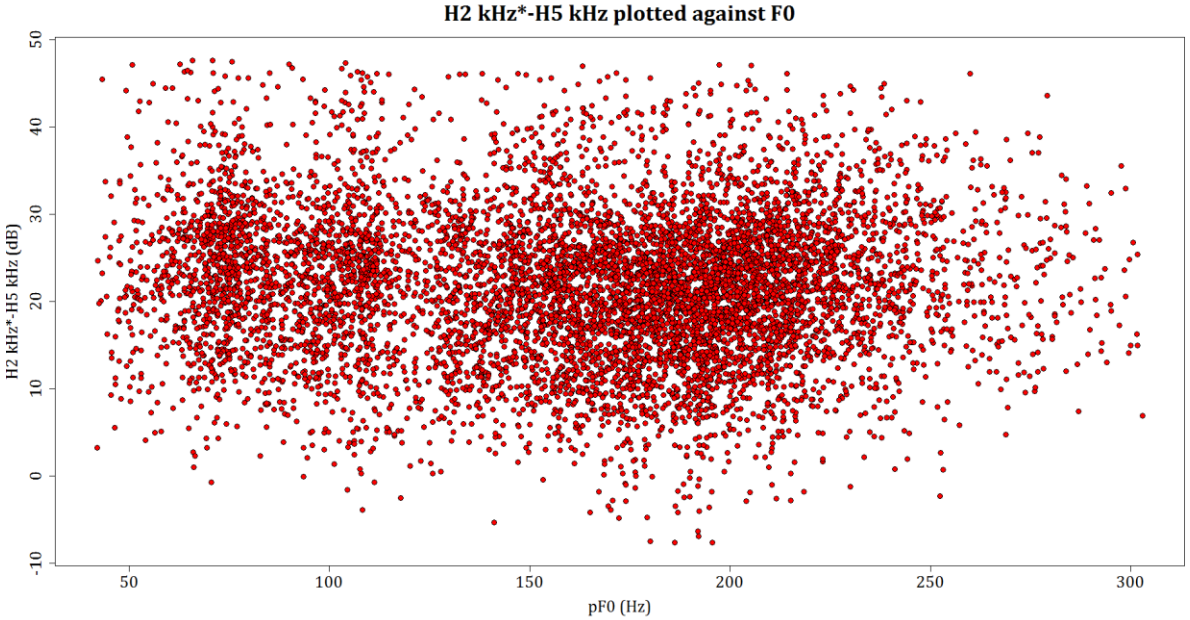

Figure E1: Scatterplot of 2K*−5K data for the whole sample plotted against pF0.

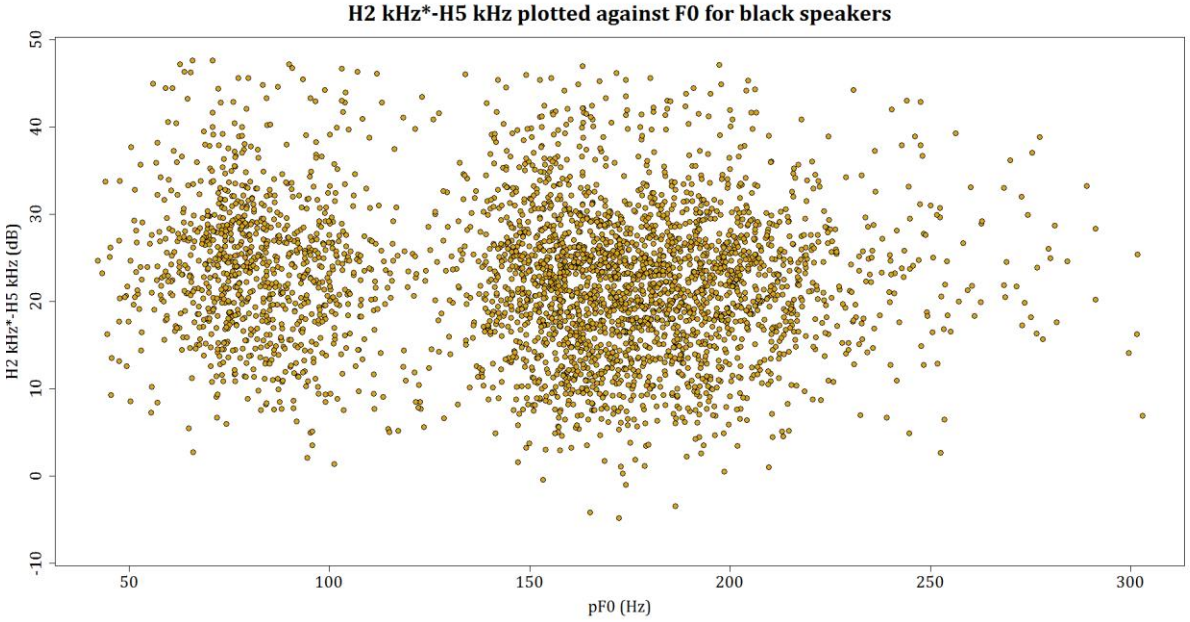

Figure E2: Scatterplot of the 2K*−5K data plotted against pF0 for the black speakers.

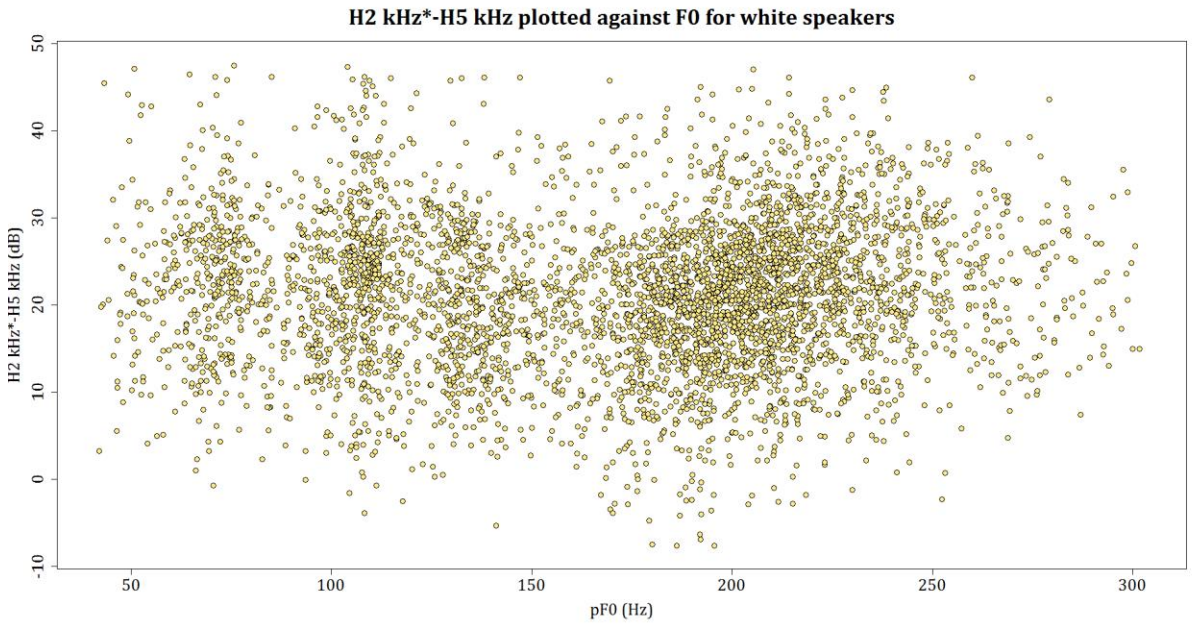

Figure E3: Scatterplot of the 2K*−5K data plotted against pF0 for the white speakers.

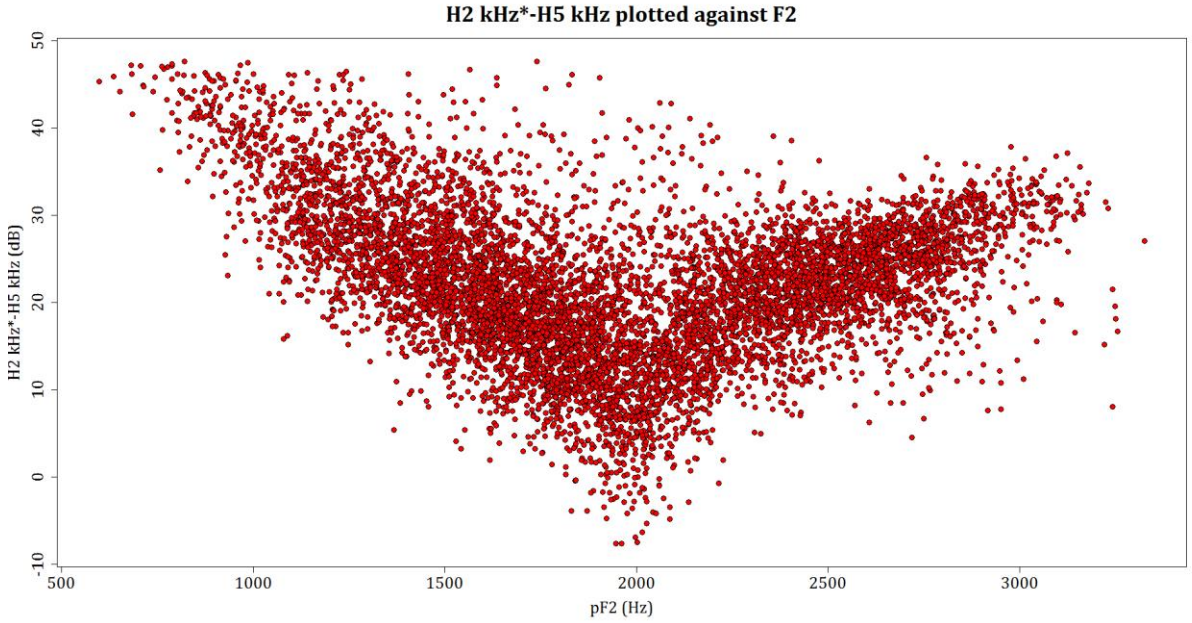

Figure E4: Scatterplot of 2K*−5K data for the whole sample plotted against pF2 measured in Hertz.

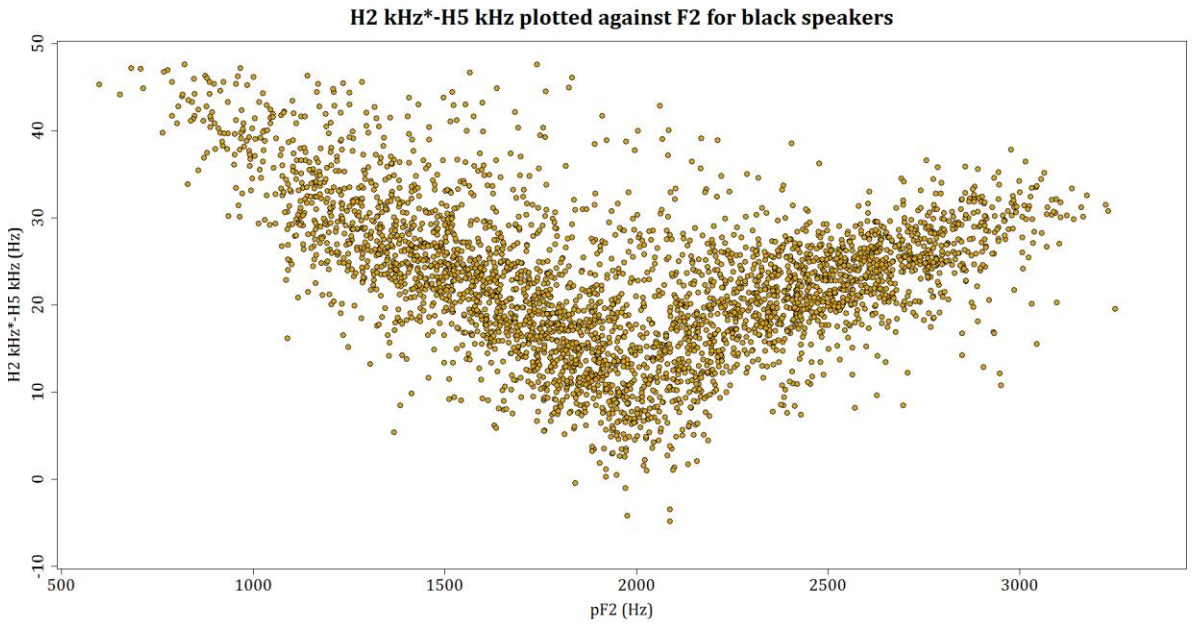

Figure E5: Scatterplot of 2K*−5K data for the black speakers plotted against pF2 measured in Hertz.

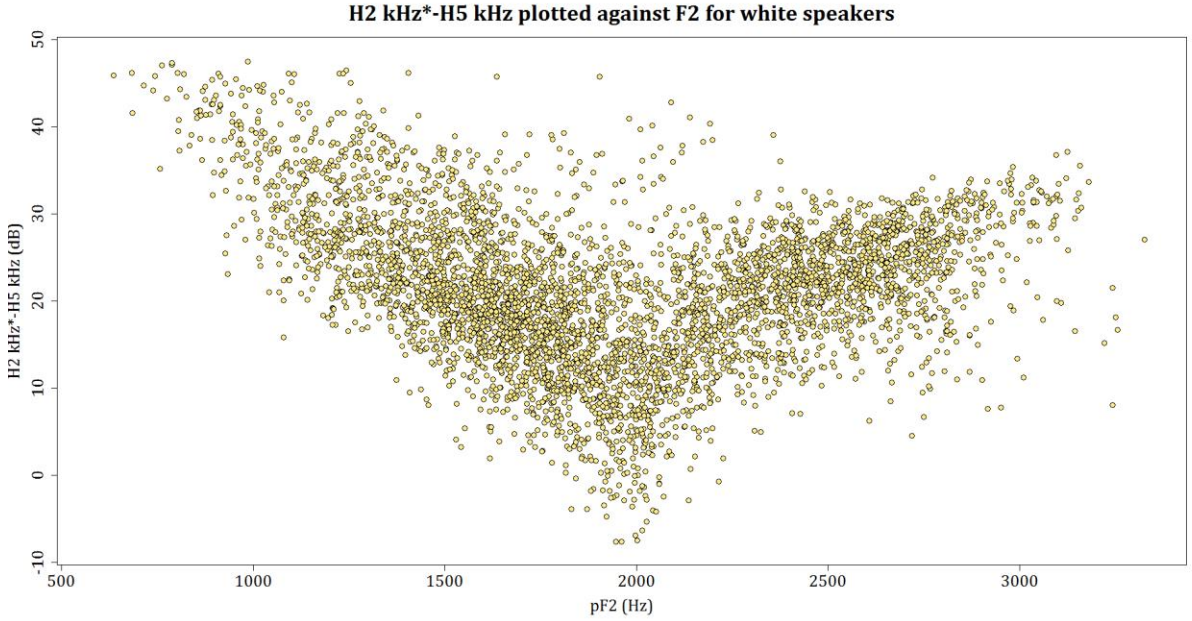

Figure E6: Scatterplot of 2K*−5K data for white speakers plotted against pF2 measured in Hertz.

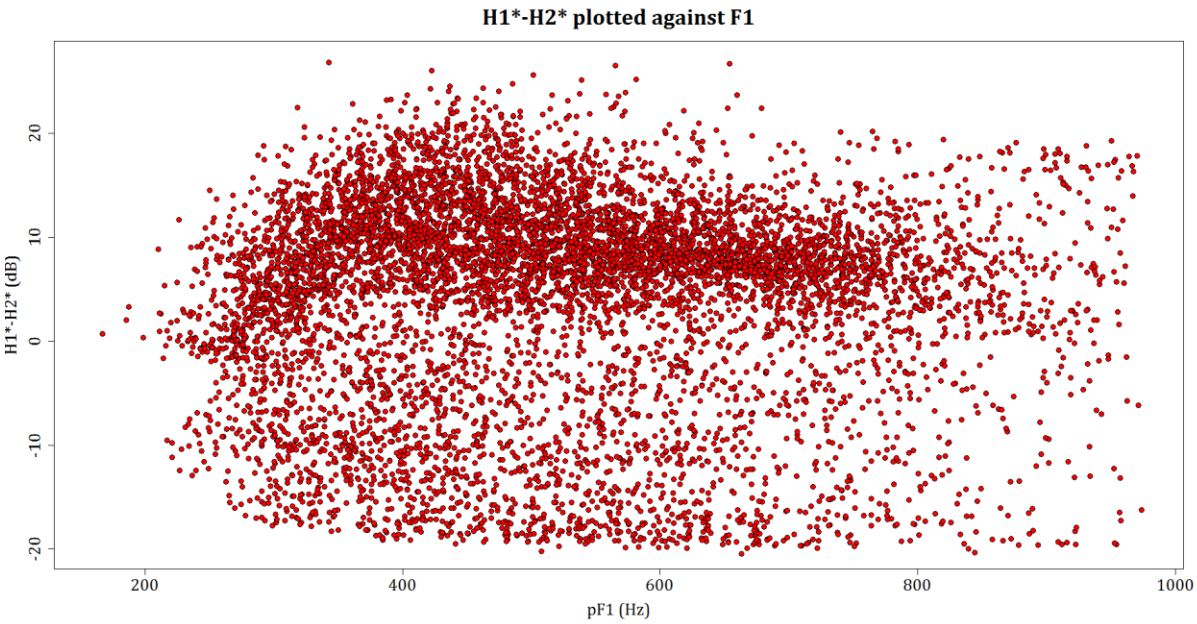

Figure E7: Scatterplot for H1*−H2* data for the interview data for the whole sample plotted against pF1 in Hertz.

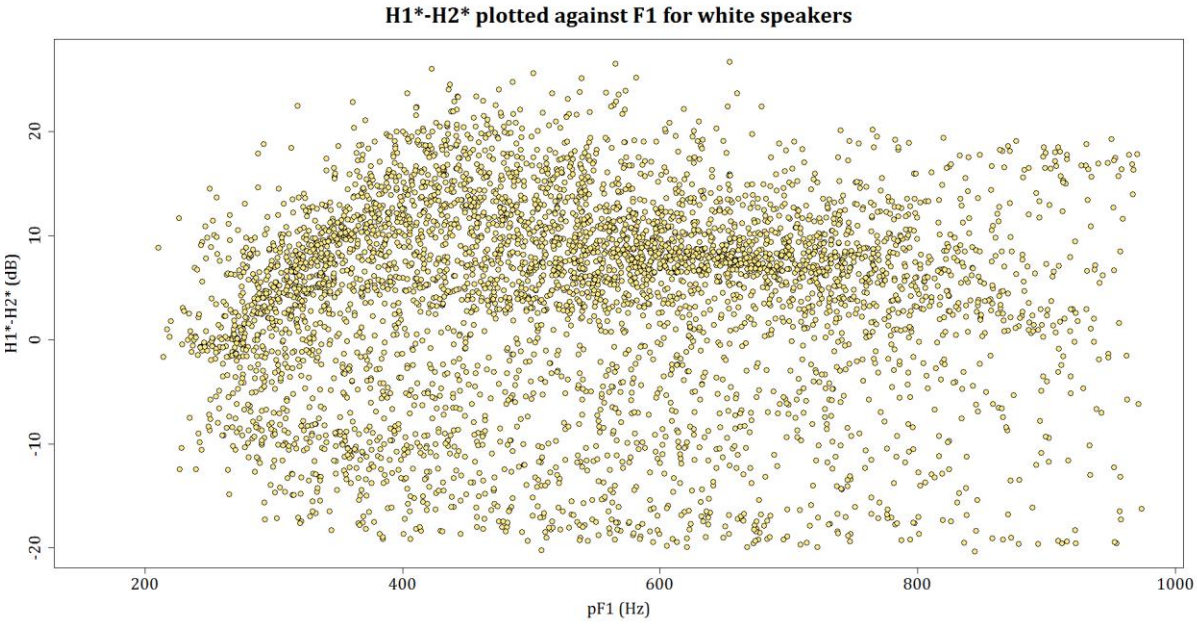

Figure E8: Scatterplot for H1*−H2* data for the white speakers plotted against pF1 in Hertz.

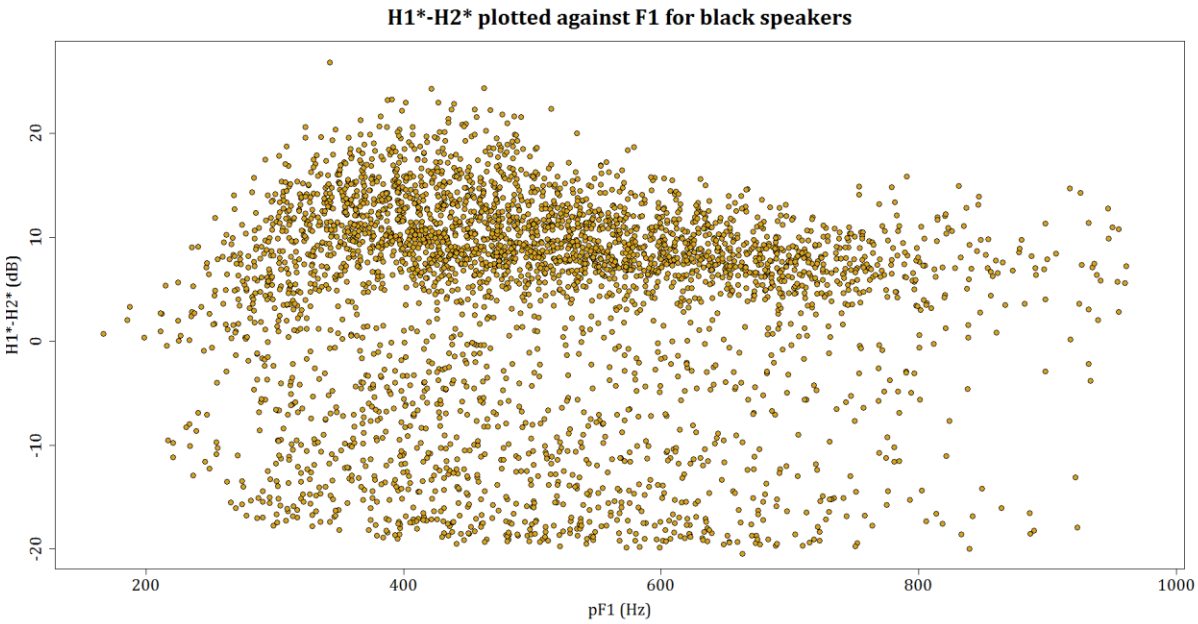

Figure E9: Scatterplot for H1*–H2* data for the black speakers plotted against pF1 in Hertz.

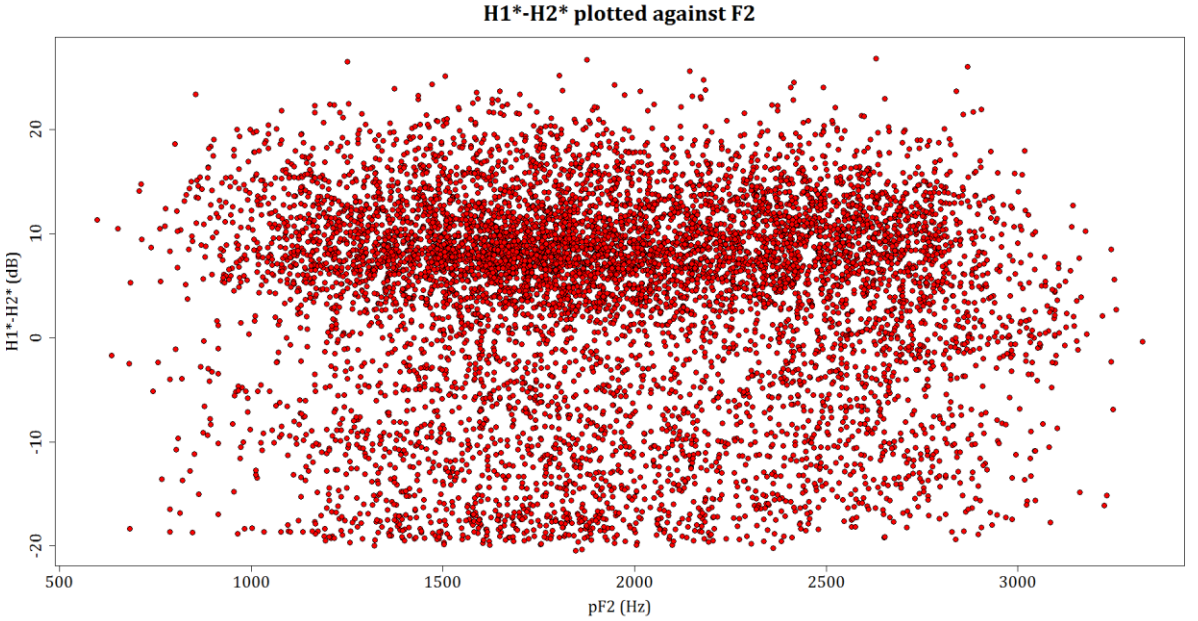

Figure E10: Scatterplot of H1*–H2* data for the whole sample plotted against pF2 in Hertz.

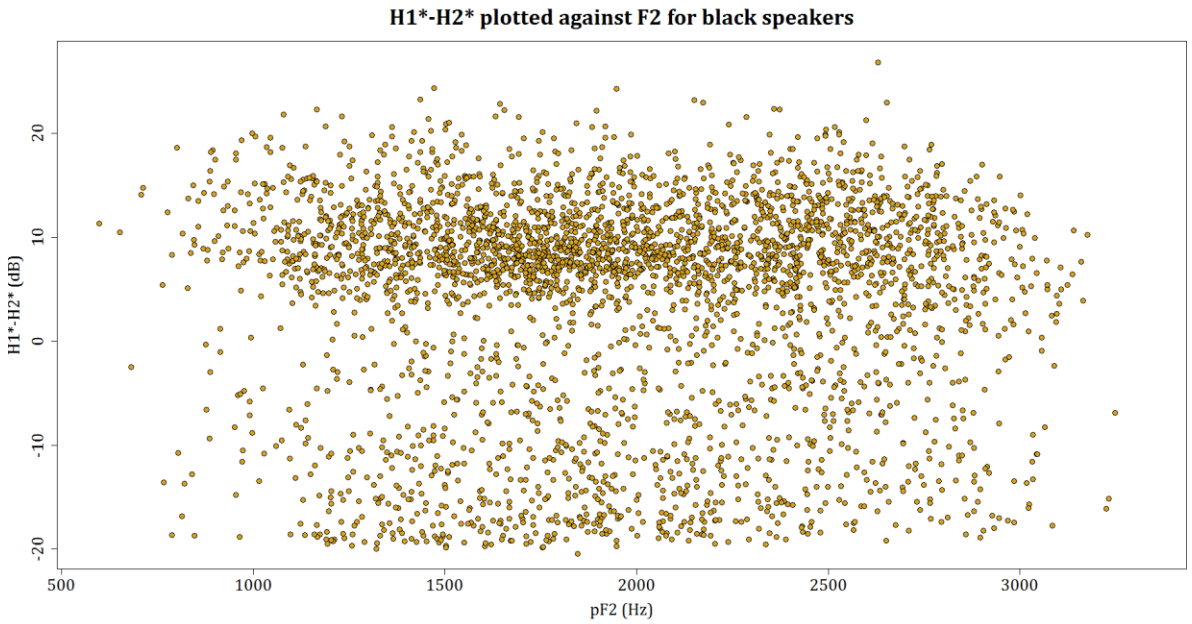

Figure E11: Scatterplot of H1*−H2* data for the black speakers plotted against pF2 in Hertz.

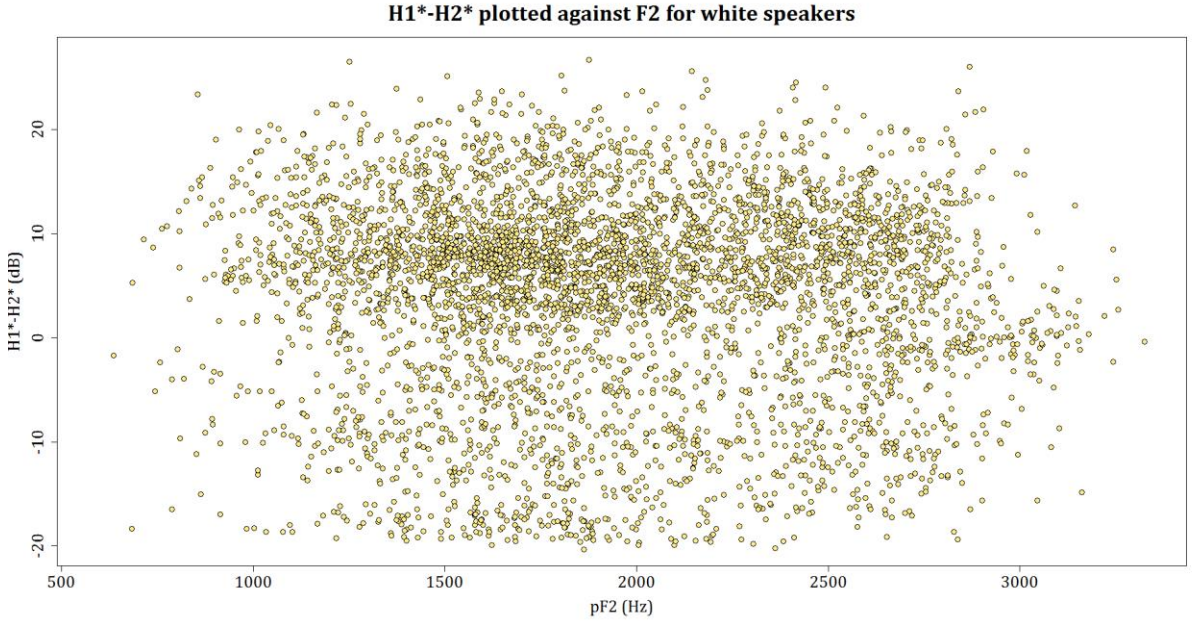

Figure E12: Scatterplot of H1*−H2* data for the white speakers plotted against pF2 in Hertz.

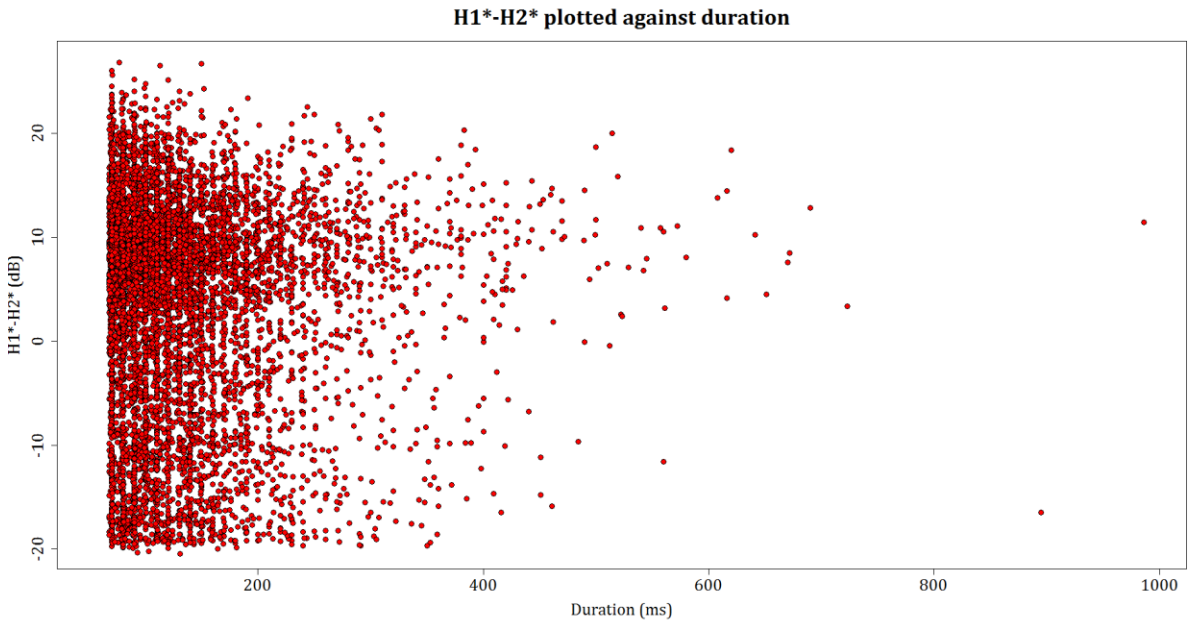

Figure E13: Scatterplot of H1*−H2* data for the whole sample plotted against duration in milliseconds.

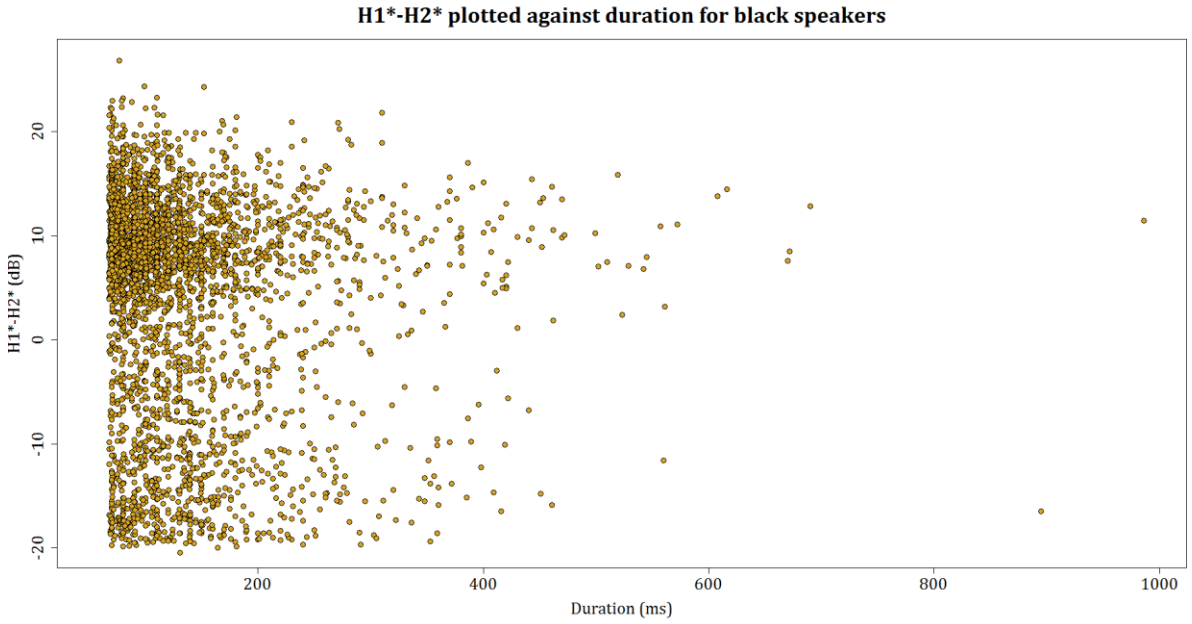

Figure E14: Scatterplot of H1*−H2* data for the black speakers plotted against duration in milliseconds.

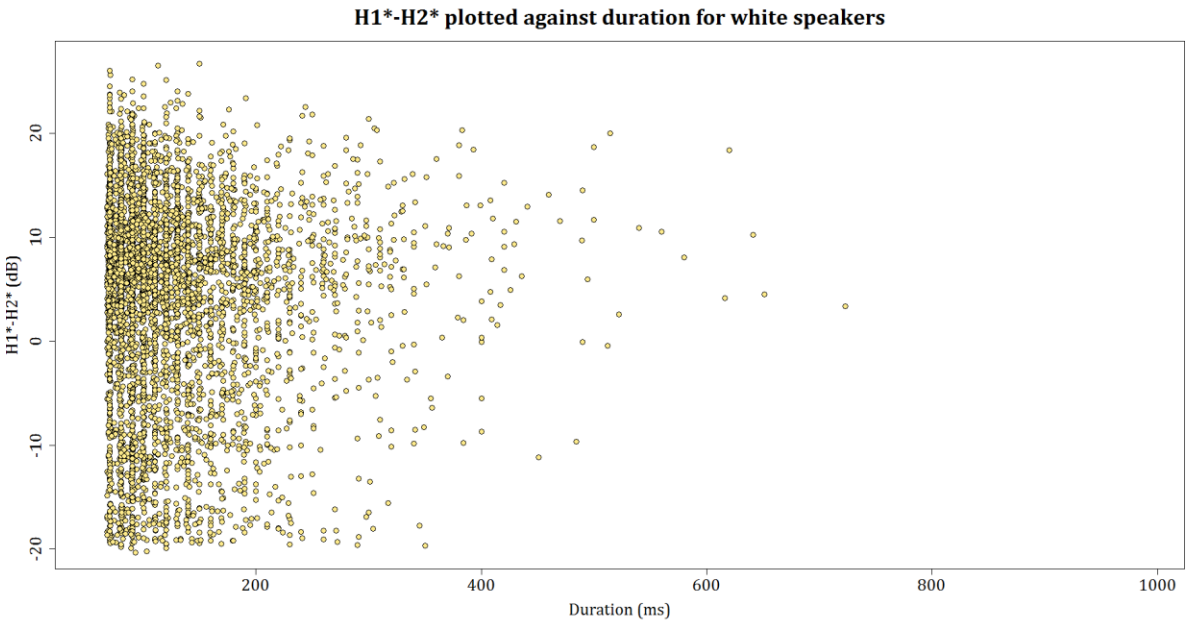

Figure E15: Scatterplot of H1*−H2* data for the white speakers plotted against duration in milliseconds.

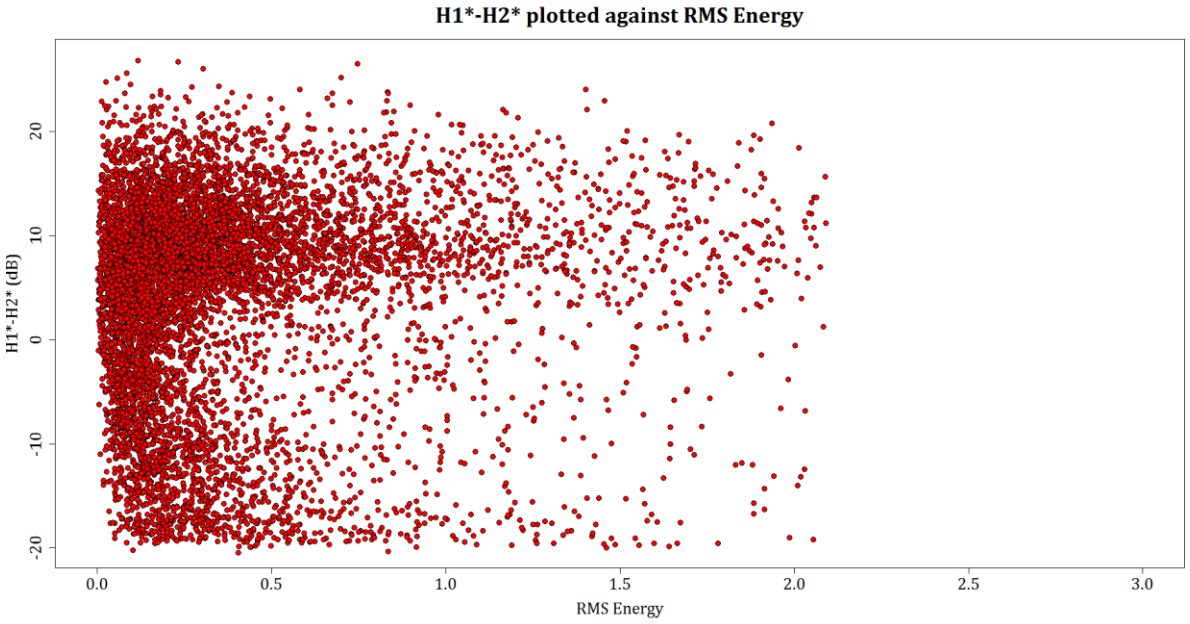

Figure E16: Scatterplot of H1*−H2* data for the whole sample plotted against RMS energy.

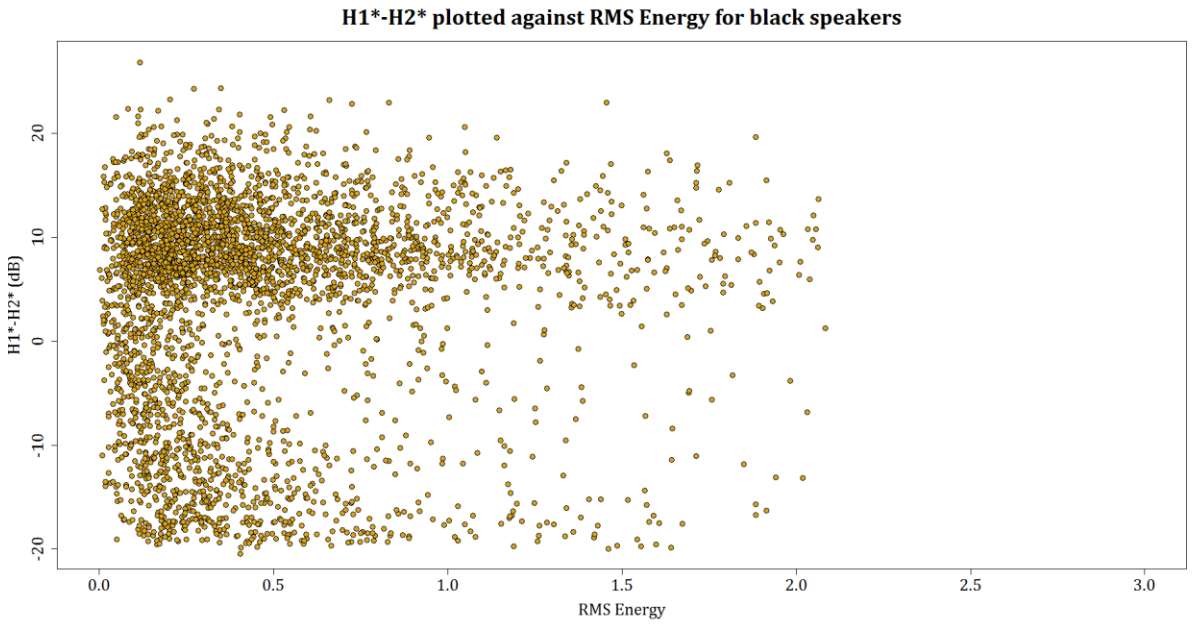

Figure E17: Scatterplot of H1*−H2* data for the black speakers plotted against RMS energy.

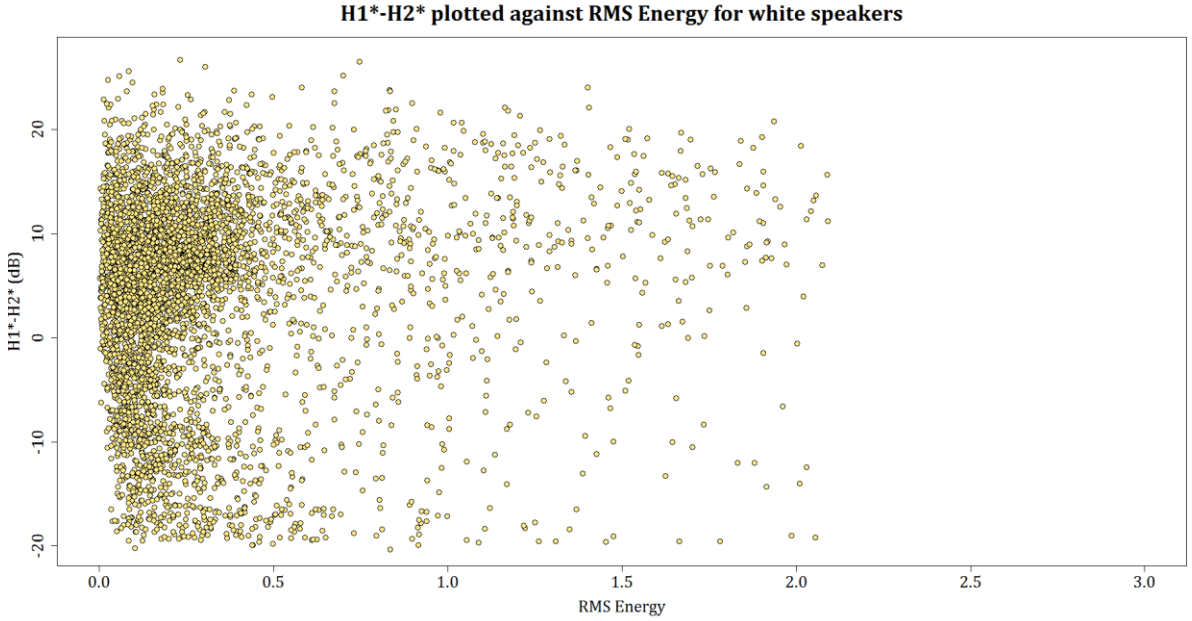

Figure E18: Scatterplot of H1*−H2* data for the white speakers plotted against RMS energy.

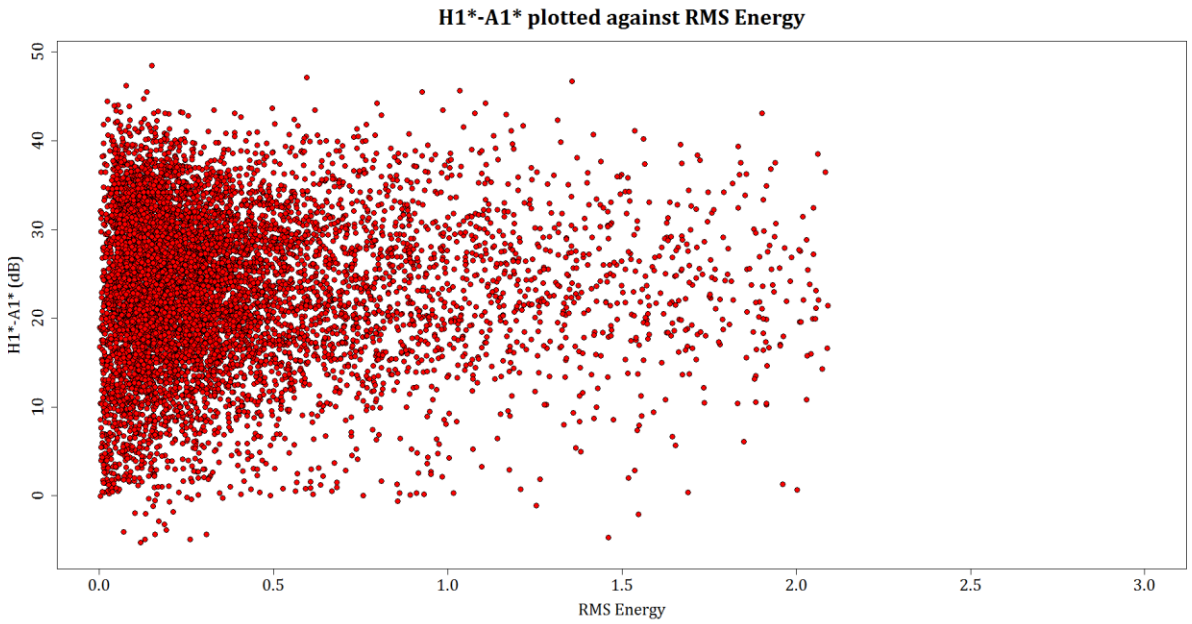

Figure E19: Scatterplot of H1*−A1* data for the whole sample plotted against RMS energy.

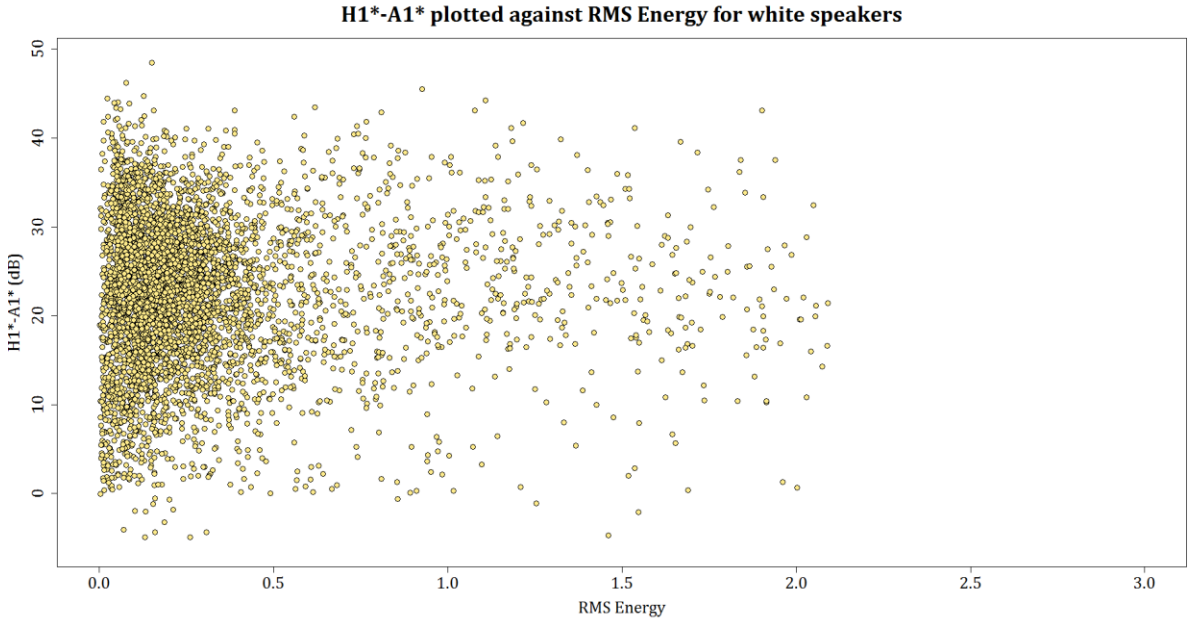

Figure E20: Scatterplot of H1*−A1* data for white speakers plotted against RMS energy.

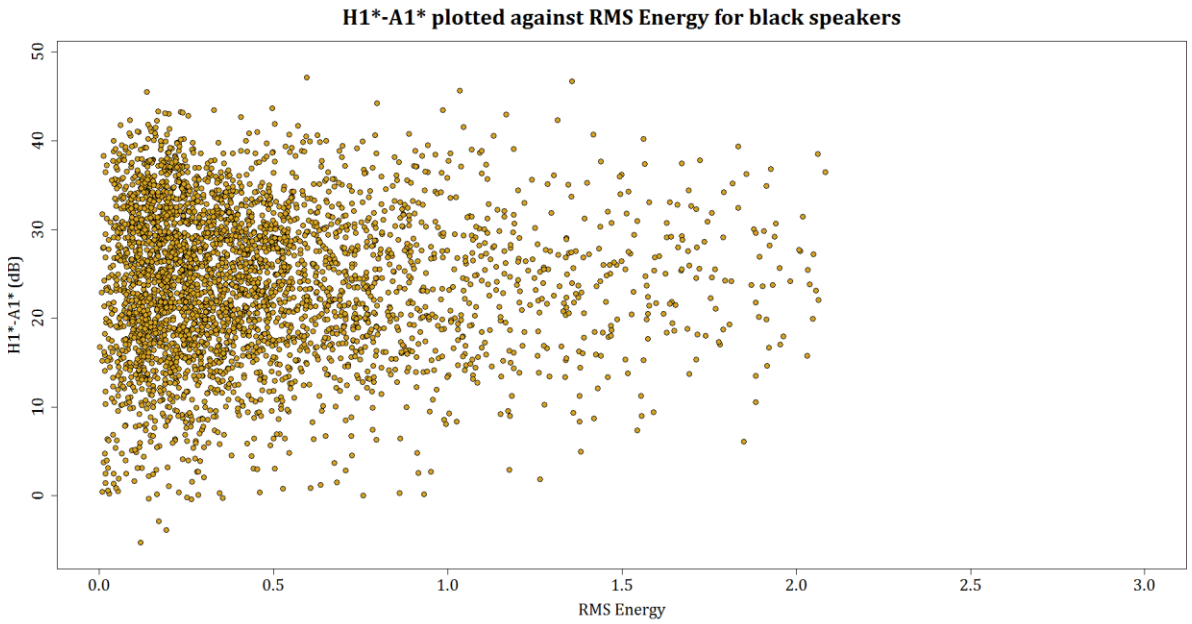

Figure E21: Scatterplot of H1*−A1* data for black speakers plotted against RMS energy.

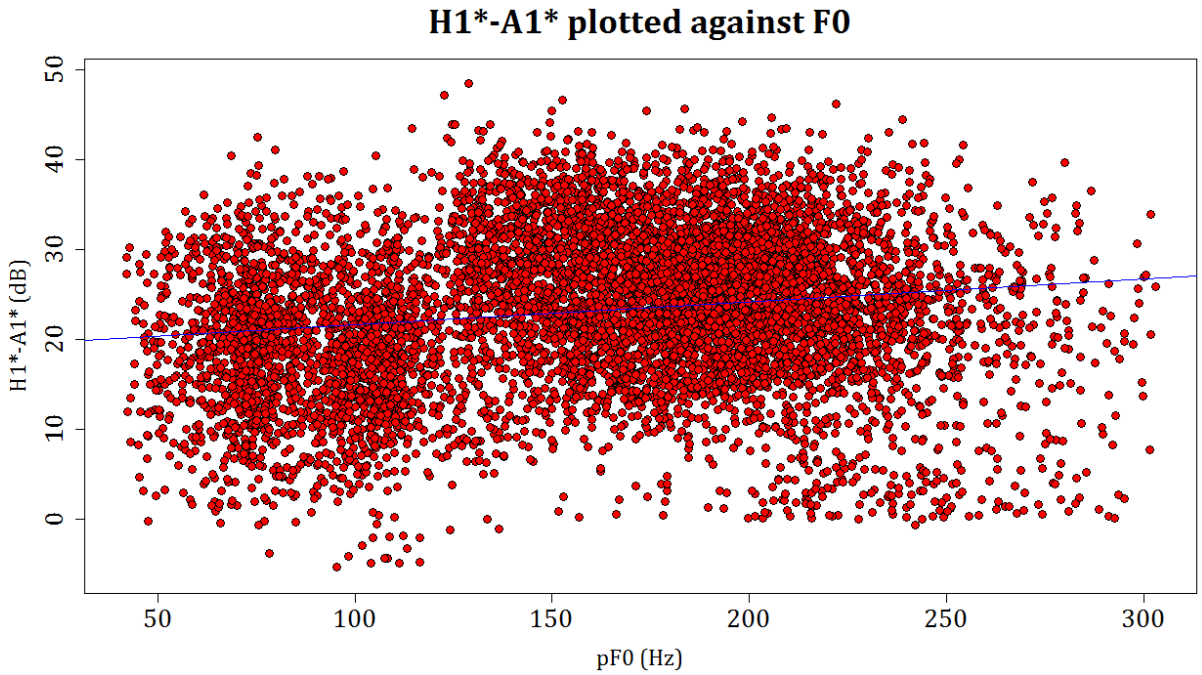

Figure E22: Scatterplot of H1*−A1* data for the whole sample plotted against pF0 in Hertz (Pearson's *r*= 0.164).

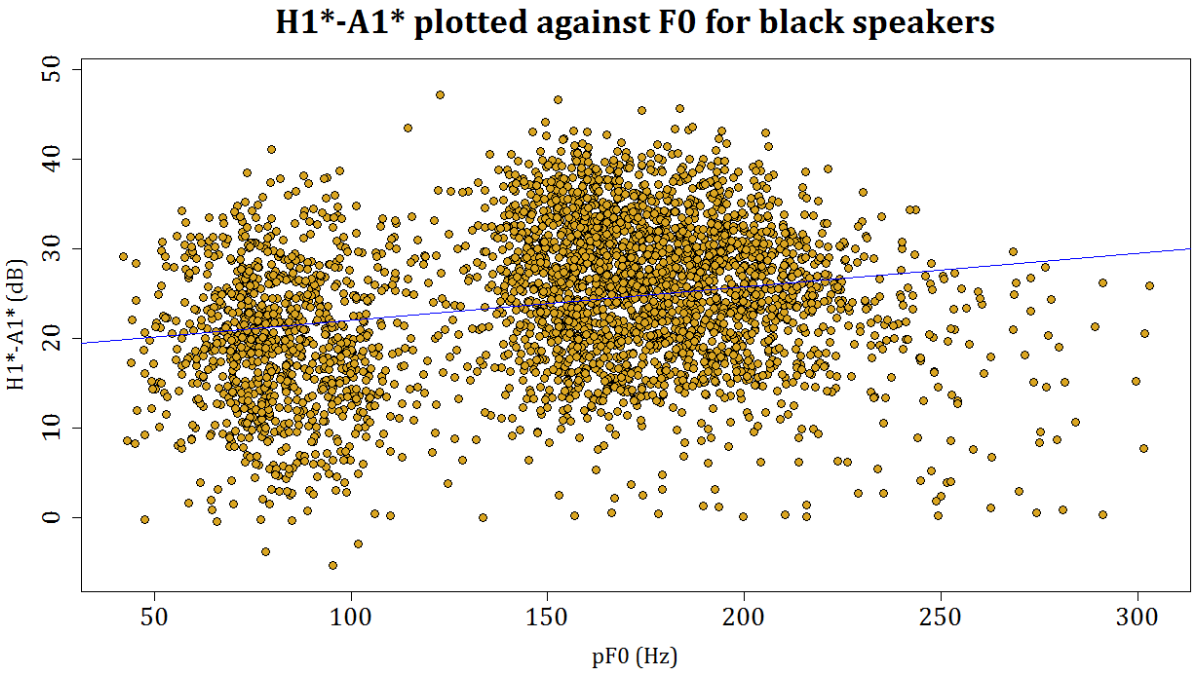

Figure E23: Scatterplot of H1*−A1* data for black speakers plotted against pF0 in Hertz (Pearson's *r*= 0.221).

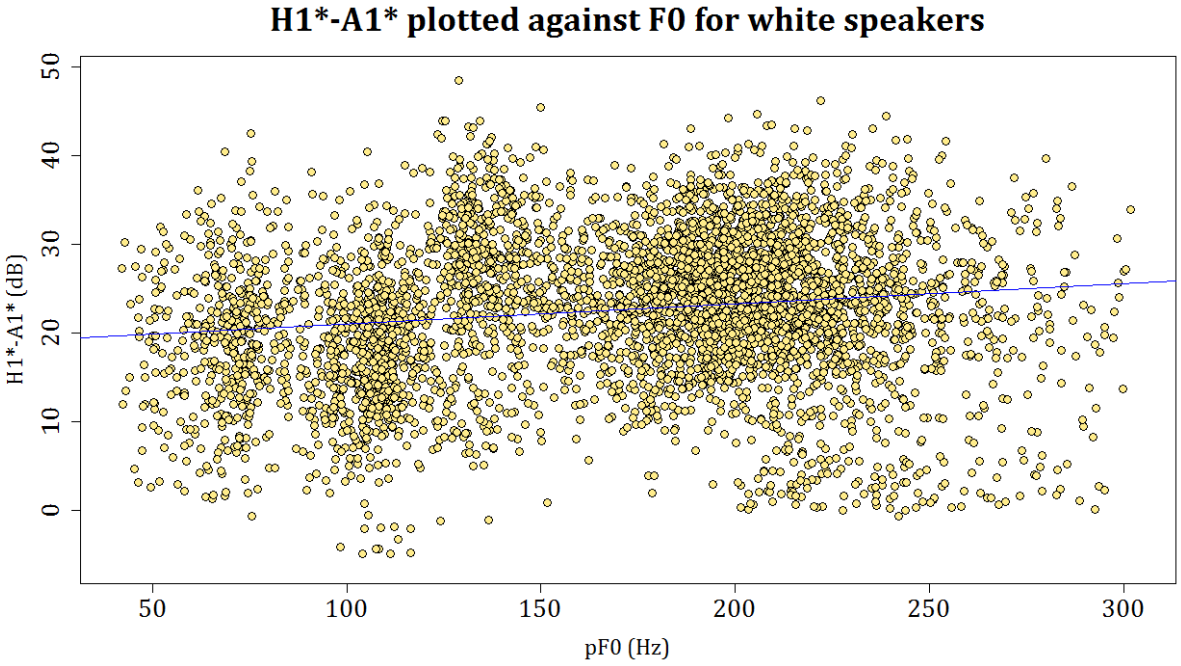

Figure E24: Scatterplot of H1*−A1* data for white speakers plotted against pF0 in Hertz (Pearson's *r*= 0.152).

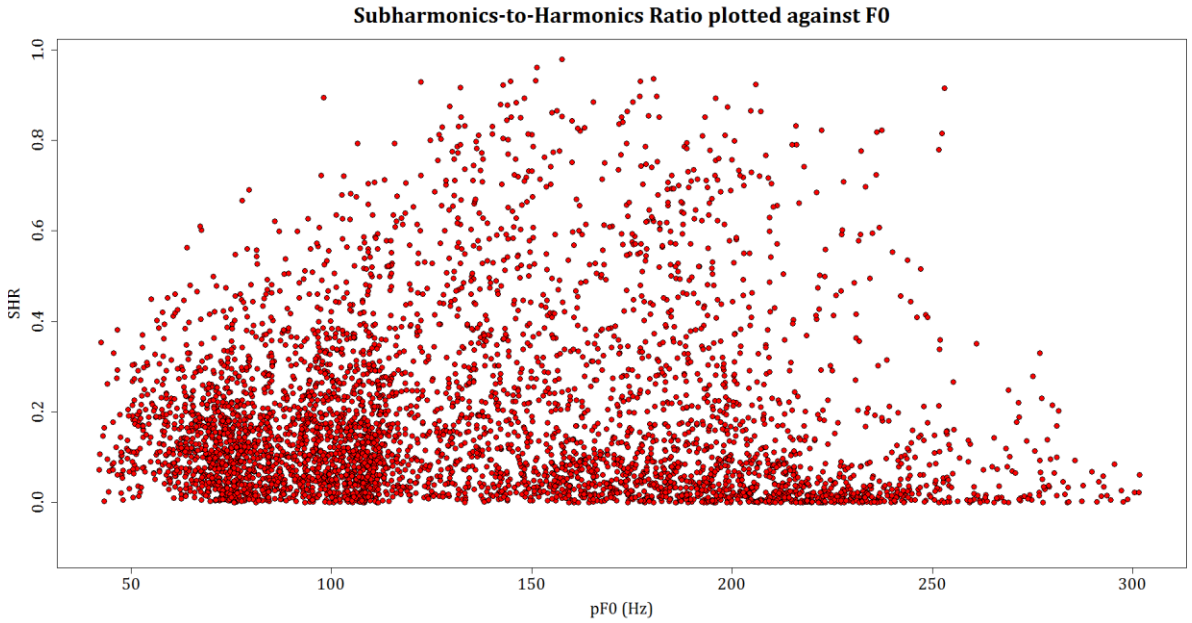

Figure E25: Scatterplot of SHR data for the whole sample plotted against pF0 in Hertz.

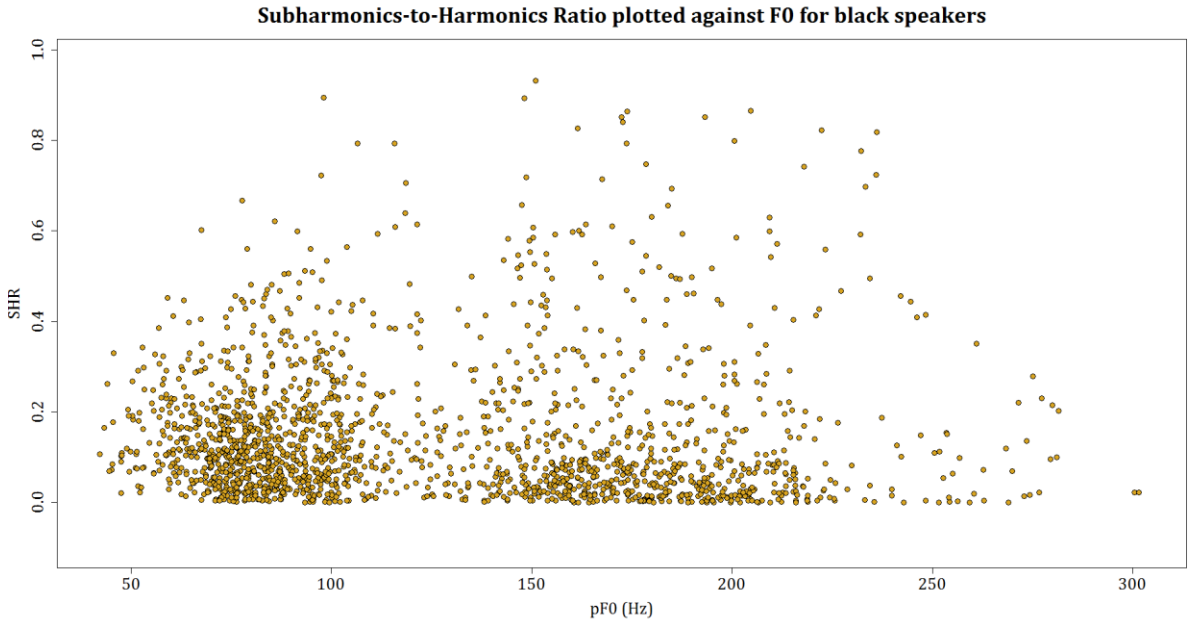

Figure E26: Scatterplot of SHR data for black speakers plotted against pF0 in Hertz.

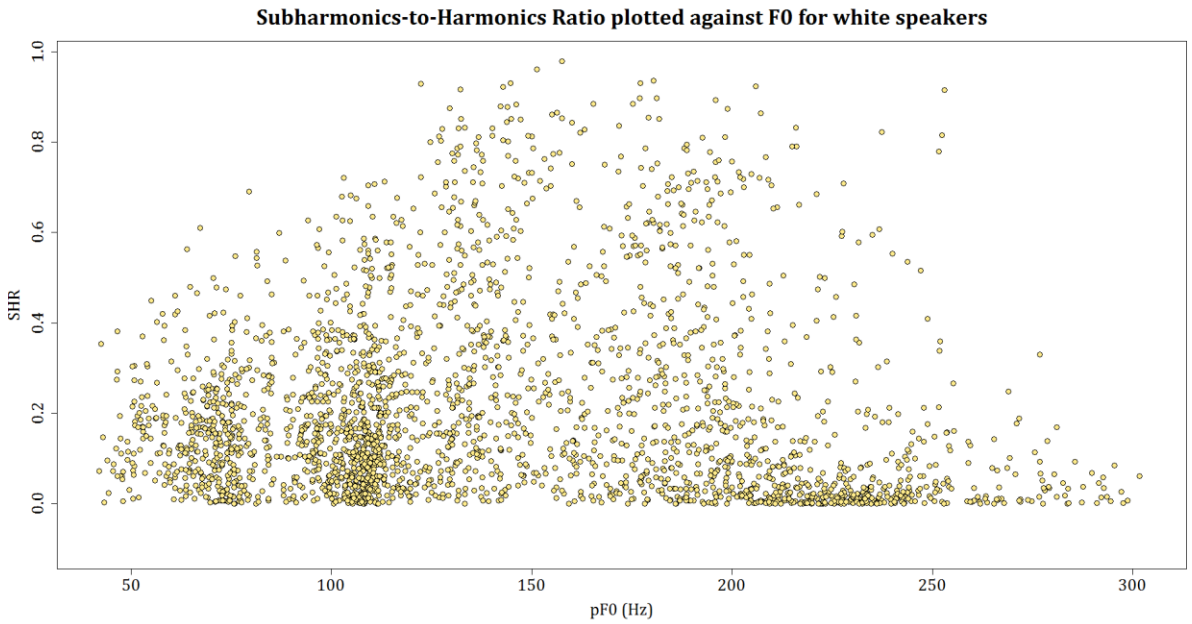

Figure E27: Scatterplot of SHR data for white speakers plotted against pF0 in Hertz.

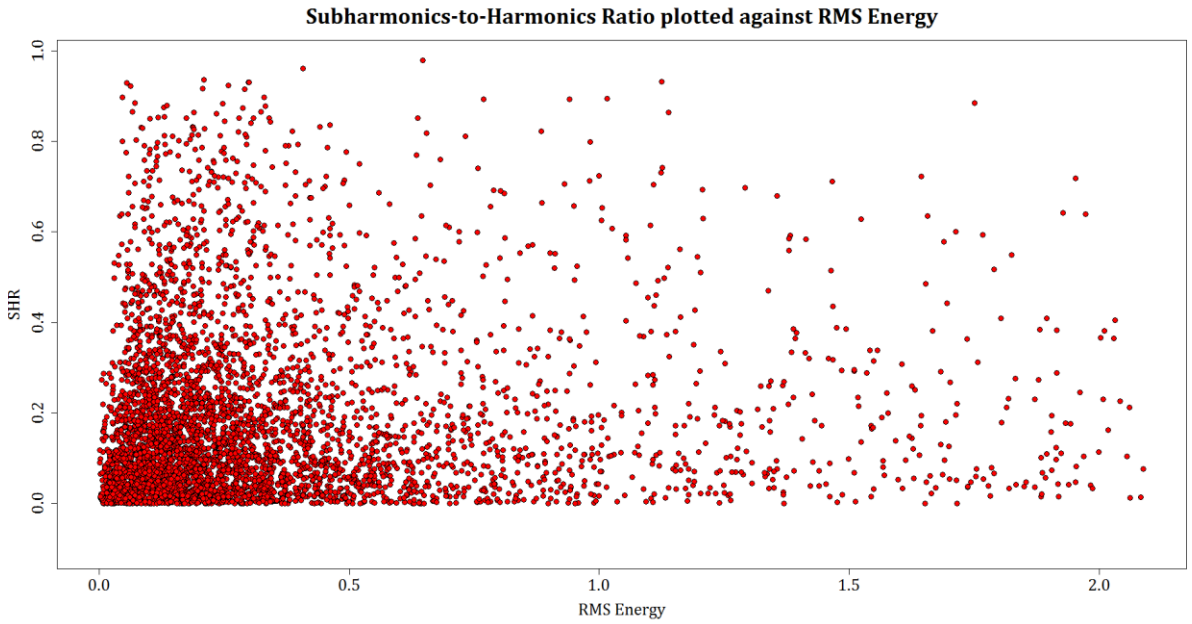

Figure E28: Scatterplot of SHR data for the whole sample plotted against RMS energy.

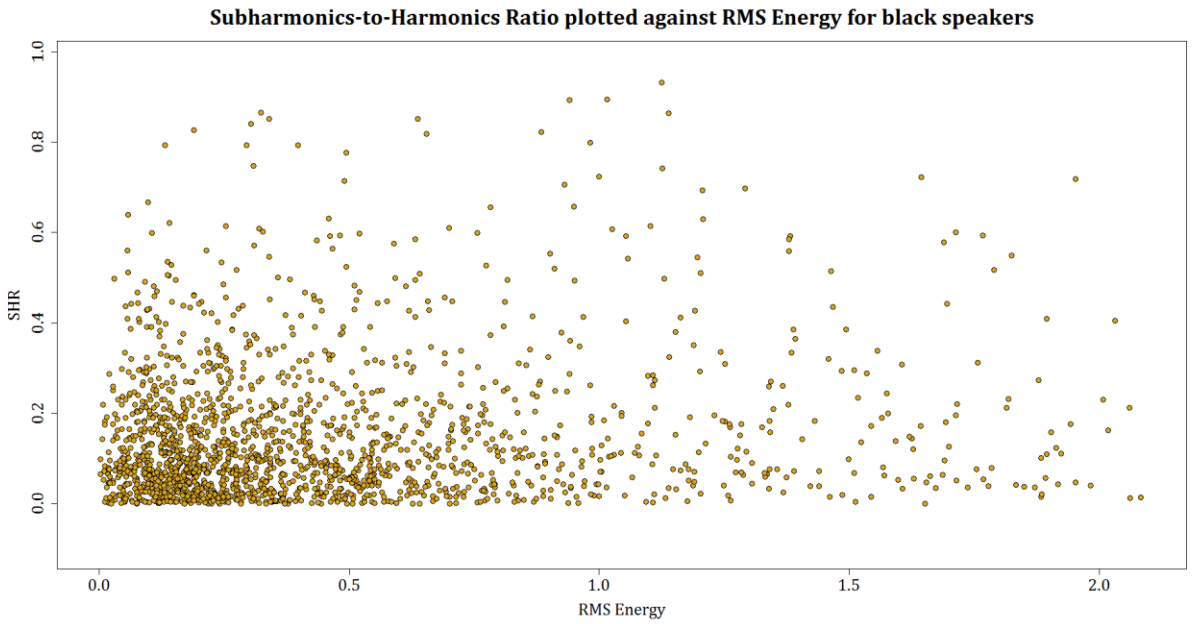

Figure E29: Scatterplot of SHR data for black speakers plotted against RMS energy.

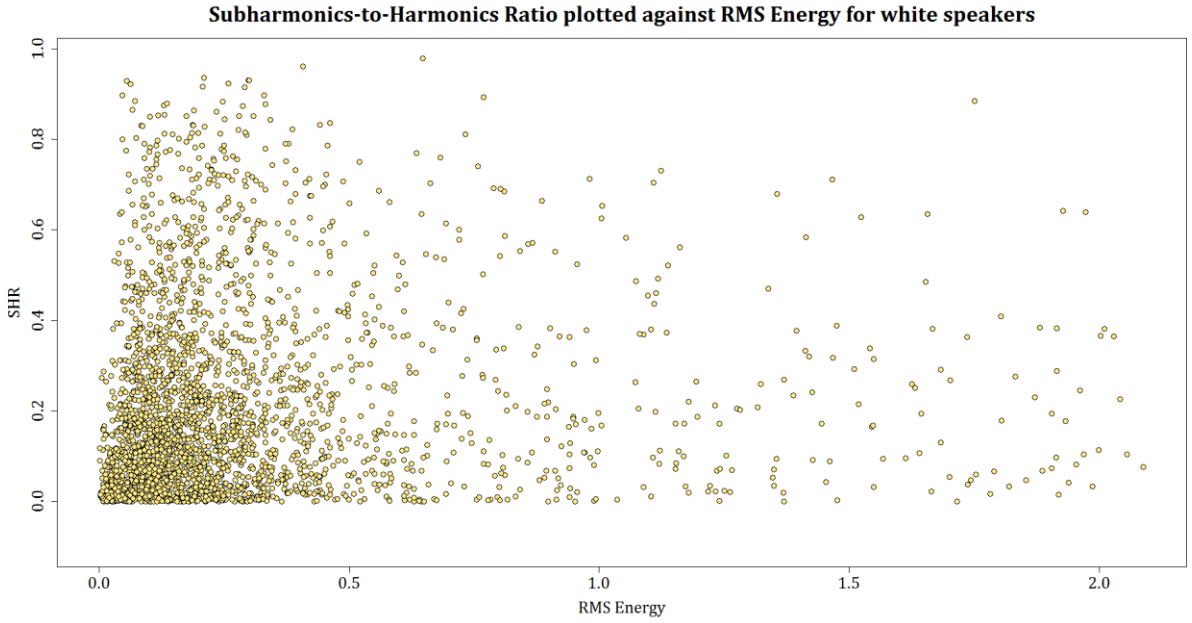

Figure E30: Scatterplot of SHR data for white speakers plotted against RMS energy.

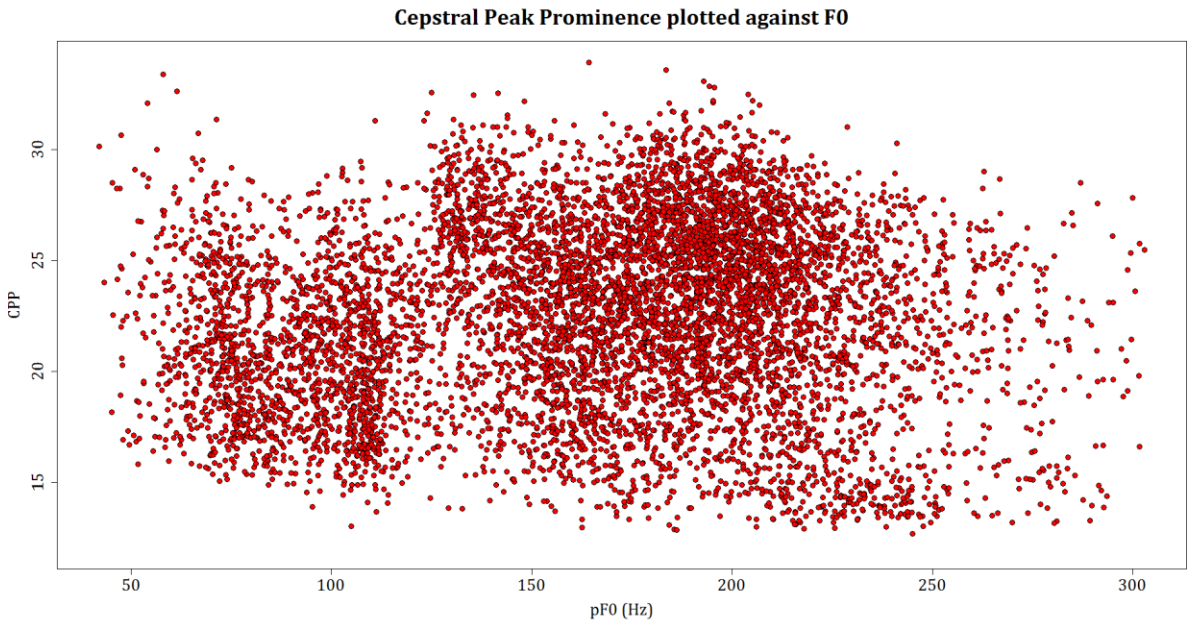

Figure E31: Scatterplot of CPP data for the whole sample plotted against pF0 in Hertz.

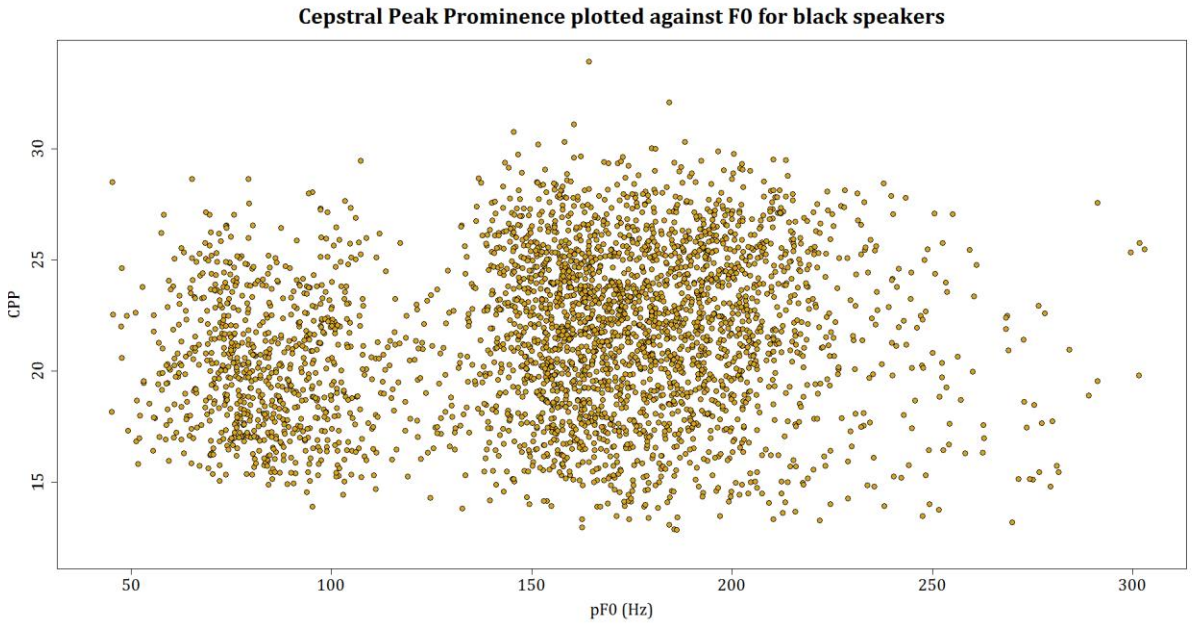

Figure E32: Scatterplot of CPP data for black speakers plotted against pF0 in Hertz.

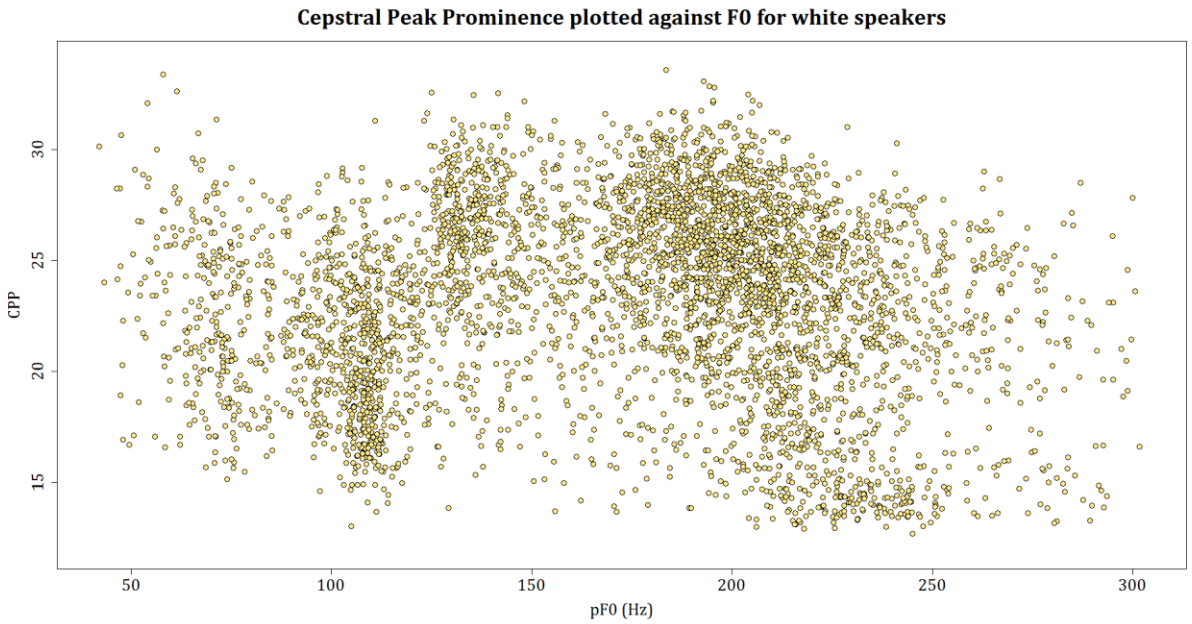

Figure E33: Scatterplot of CPP data for white speakers plotted against pF0 in Hertz.

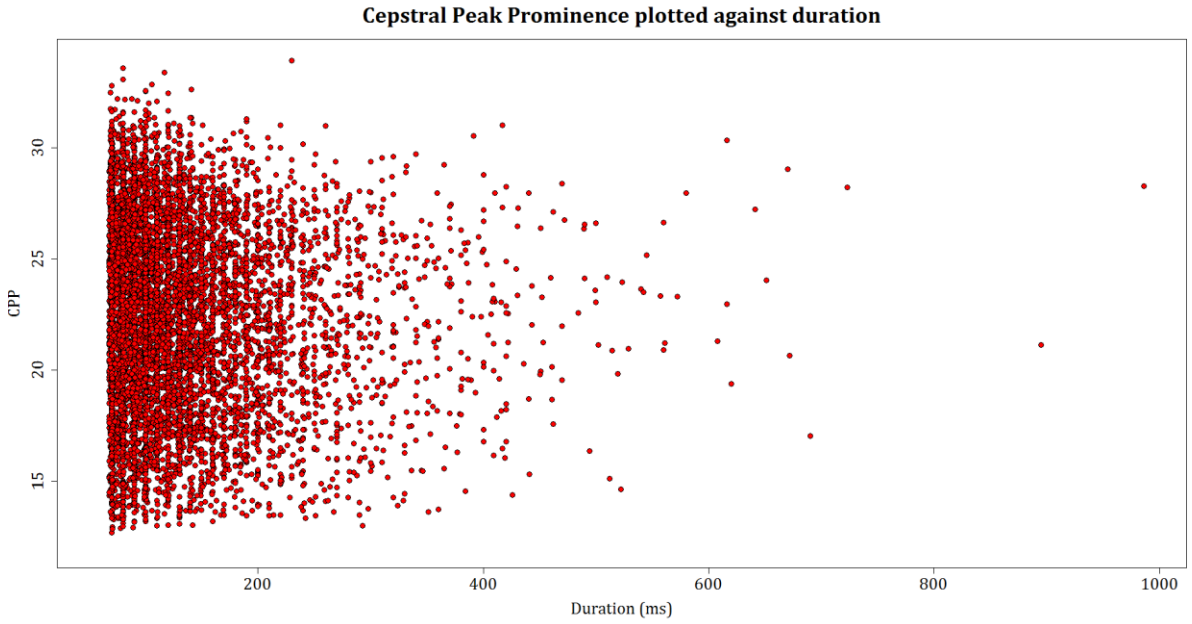

Figure E34: Scatterplot of CPP data for the whole sample plotted against duration in milliseconds.

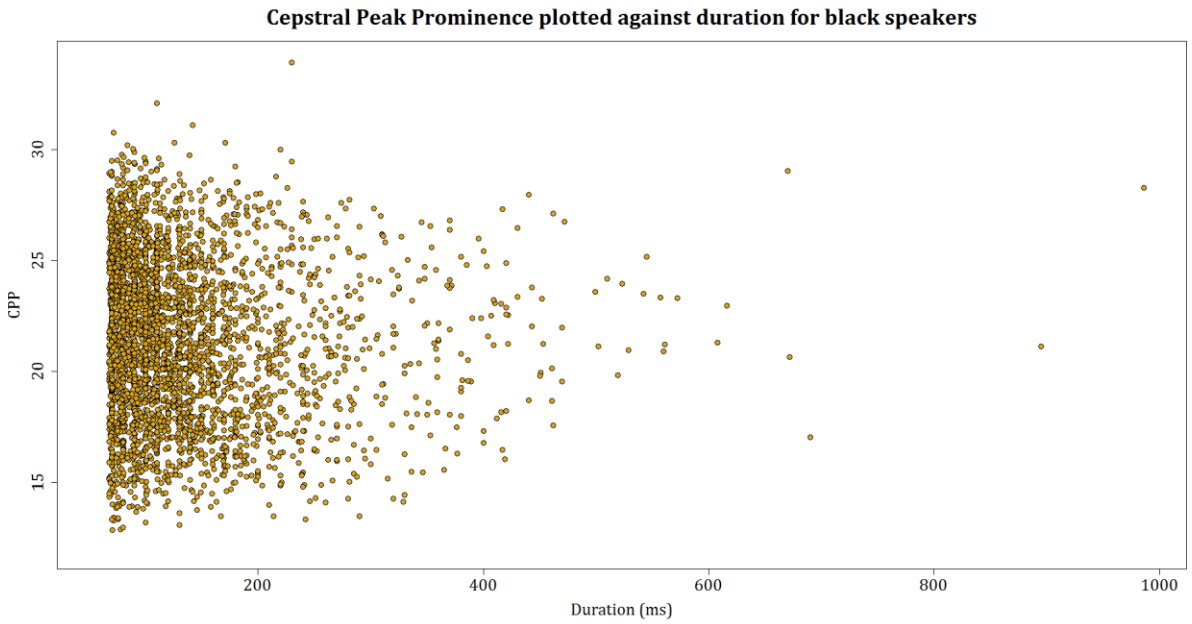

Figure E35: Scatterplot of CPP data for black speakers plotted against duration in milliseconds.

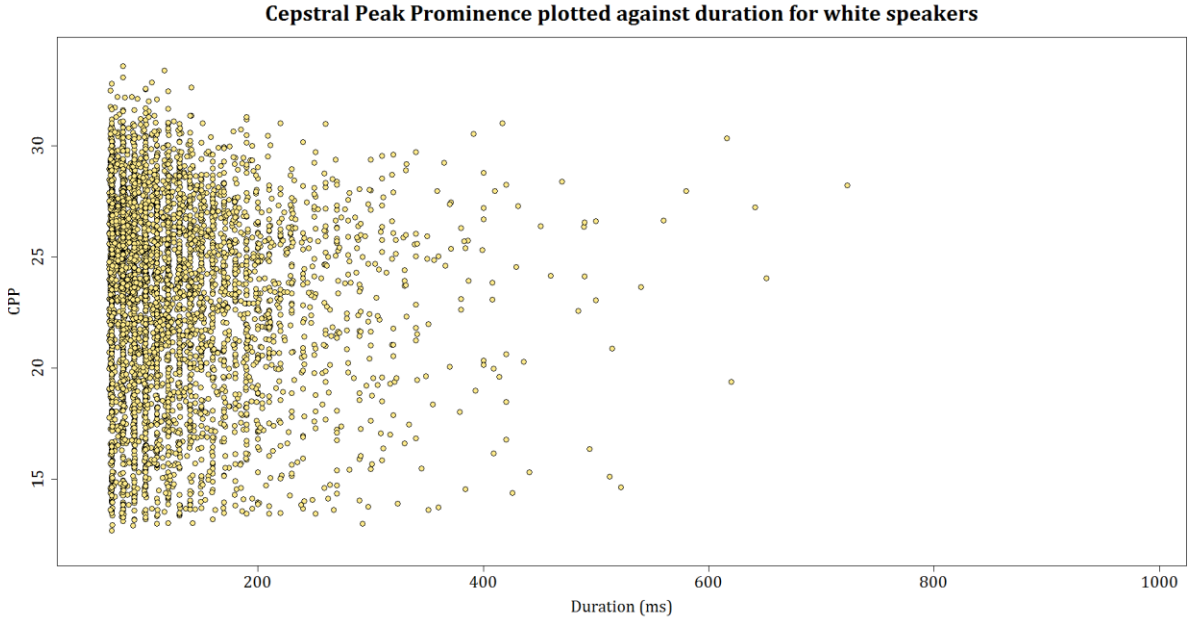

Figure E36: Scatterplot of CPP data for white speakers plotted against duration in milliseconds.

References:

Abercrombie, David. 1967. *Elements of General Phonetics*. Chicago: Aldine.

Addington, D. W. 1963. *The relationship of certain vocal characteristics with perceived speaker characteristics*. PhD dissertation. State University of Iowa.

Ajmani, M. L. 1990. A metrical study of the laryngeal skeleton in adult Nigerians. *Journal of Anatomy 171*:187-191.

Alim, H. S. 2004. *You Know My Steez: An Ethnographic and Sociolinguistic Study of Styleshifting in a Black American Speech Community*. Durham, NC: Duke University Press.

Altenberg, E.P. and C.T. Ferrand. 2006. Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. *Journal of Voice 20*:98-96.

Andrianopoulos, M.V., K. Darrow and J. Chen. 2001. Multimodal standardization of voice among four multicultural populations: formant structures. *Journal of Voice 15*:61-77.

Andruski, J. 2006. Tone clarity in mixed pitch/phonation-type tones. *Journal of Phonetics 34*: 388-404.

Ashby, M., J. Przedlacka. 2014. Measuring incompleteness: Acoustic correlates of glottal articulation. *Journal of the International Phonetic Association 44*: 283-296.

Awan, Shaheen N. and P.B. Mueller. 1996. Speaking fundamental frequency characteristics of white, African American, and Hispanic kindergartners. *Journal of Speech and Hearing research 39*:573-657.

Awan, Shaheen N., Ashley Giovinco and Jennifer Owens. 2012. Effects of vocal intensity and vowel type on cepstral analysis of voice. *Journal of Voice 26*: 670.e15-670.e20.

Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Barr, D.J., C. Scheepers and H.J. Tilly. 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language 68*: 255-278.

Bates, Douglas, Martin Maechler, Ben Bolker and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software 67(1)*: 1-48. <doi:10.18637/jss.v067.i01>.

Bekker, Ian. 2009. *The Vowels of South African English*. Unpublished PhD thesis. Potchefstroom: North-West University.

Berry, Donald A. 1987. Logarithmic transformations in ANOVA. *Biometrics 43(2)*: 439-456.

Bickley, C. 1982. Acoustic analysis and perception of breathy vowels. In *Speech Communication Group Working Papers*. 71-82. Cambridge: Massachusetts Institute of Technology.

Bishop, J. and P. Keating. 2012. Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *Journal of the Acoustical Society of America 132*:1100-1112.

Blankenship, Barbara 2002. The timing of nonmodal phonation in vowels. *Journal of Phonetics 30*:163–191.

Blankenship, Barbara. 1997. *The Time Course of Breathiness and Laryngealization in Vowels*. Ph.D. dissertation, University of California, Los Angeles.

Boersma, P. and D. Weeknink. 2015. 'Praat: Doing phonetics by computer (version 5.4.17)'. Computer Program: Retrieved August 30, 2015, from http://www.praat.org.

Boshoff, P.H. 1945. The anatomy of the South African Negro larynx. *South African Journal of Medical Sciences 10*:113-119.

Britt, E. 2011. Can the church say amen: Strategic uses of black preaching style at the State of the Black Union. *Language in Society 40(2)*:211–233. doi: http://dx.doi.org/10.1017/S0047404511000042

Bucholtz, M. and K. Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies 7(4–5)*: 585–614.

Catford, John C. 1964. Phonation types: The classification of some laryngeal components of speech production. In David Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott, and J.L.M. Trim (eds.) *In Honour of Daniel Jones*. London: Longmans. 26-37.

Chen, G., S. J. Park, J. Kreiman and A. Alwan. 2014. Investigating the effect of F0 and vocal intensity on harmonic magnitudes: Data from high-speed laryngeal videoendoscopy. Paper presented at the Fifteenth Annual Conference of the International Speech and Communication Association.

Clark, John and Colin Yallop. 1990. The Acoustics of speech production in *An Introduction to Phonetics and Phonology*. Oxford, UK: Cambridge, Massachusetts, USA: Blackwell. 199-200.

CMUdict 2015. The CMU Pronouncing Dictionary. Website. Accessible at http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Crystal, David. 1969. *Prosodic Systems and Intonation in English*. C.U.P.

de Krom, G. 1993. A cepstrum-based technique for determining a harmonic-to-noise ratio in speech signals. *Journal of Speech and Hearing Research 36*: 254-66.

Denes, P. B. and Pinson, E. N. 1993. *The Speech Chain: The physics and biology of spoken language, 2nd edition*. New York: W. H. Freeman.

Deumert, Ana. 2013. Xhosa in town (revisited) – space, place and language. *International Journal of the Sociology of Language 222*: 51-75.

DiCanio, Christian T. 2009. The phonetics of register in Takhian Thong Chong. *Journal of the International Phonetic Association 39*: 162–88.

Dickens, M. and G. Sawyer. 1962. An experimental comparison of vocal quality among mixed groups of Whites and Negroes. *Southern Speech Journal 18*:178-185.

Dillard, J.L. 1972. *Black English: Its History and Usage in the United States*. New York: Vintage.

Edmondson, Jerold. A. and John Henry Esling. 2006. The valves of the throat and their functioning in tone. *Vocal register and stress: Laryngoscopic case studies, Phonology 23*: 157-91.

Esling, John Henry and J. G. Harris. 2005. States of the glottis: An articulatory phonetic model based on laryngoscopic observations. In William. J. Hardcastle and J.M. Beck (eds.) *A Figure of Speech: A Festschrift for John Laver*. Mahwah, New Jersey: Lawrence Erlbaum Associates. 347-83.

Esling, John Henry and Jerold A. Edmonson. 2011. Acoustical analysis of voice quality for sociophonetic purposes. In Marrianna Di Paolo, Malcah Yaeger-Dror, and Alicia Beckford Wassink (eds.) *Sociophonetics: A Student's Guide*. London, New York: Routledge. 131-148.

Esling, John Henry. 1978. *Voice quality in Edinburgh: A Sociolinguistic and Phonetic Study*, Ph.D. dissertation: University of Edinburgh.

Esling, John Henry. 2006. Voice quality. In R. E. Asher and J. M. Y. Simpson (eds.) *The Encyclopedia of Language and Linguistics, 2nd ed*. Oxford: Pergamon Press.

Esling, John Henry. 2010. Phonetic notation. In William J. Hardcastle. John Laver, and F. E. Gibbon (eds.) *Handbook of Phonetic Sciences, 2nd ed*. Oxford: Wiley-Blackwell. 678-702.

Esposito, C.M. 2006. *The Effects of Linguistic Experience on the Perception of Phonation*. Ph.D. dissertation: UCLA.

Esposito, C.M. 2010. The effects of linguistic experience on the perception of phonation. *Journal of Phonetics 38*:306-316.

Esposito, C.M.. 2010. Variation in contrastive phonation in Santa Ana Del Valle Zapotec. *Journal of the International Phonetic Association 40*: 181-198.

Everitt, Brian S. and Torsten Hothorn. 2006. *A Handbook of Statistical Analyses Using R*. Boca Raton: Chapman & Hall/CRC Taylor & Francis Group.

Fay, P. J. and W.C. Middleton. 1939. Judgement of occupation from the voice as transmitted over a public address system and over a radio. *Journal of Applied Psychology 23*:586-601.

Finlayson, R. 1989. *An Introduction to Xhosa Phonetics*. Constantia, South Africa: M. Lubbe.

Fischer-Jørgensen, Eli. 1967. Phonetic analysis of breathy (murmured) vowels. *Journal of Indian Linguistics 28*: 71-139.

Foulkes, P. and G. Docherty. 2006. The social life of phonetics and phonology. *Journal of Phonetics 34*: 409–38.

Fraile, Rubén and Juan Ignacio Godino-Llorente. 2014. Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control 14*: 42-54.

Fung, Roxana. 2015. Voice quality: A preliminary study on the phonetic distinctions of two Cantonese accents. *Proceedings of the International Phonetic Association*, Glasgow, UK.

Gaither, Sarah E., Ariel M. Cohen-Goldberg, Calvin L. Gidney and Keith B. Maddox. 2015. Sounding Black or White: priming identity and biracial speech. *Frontiers in Psychology 6*: 457.

Garellek, M. 2012. The timing and sequencing of coarticulated non-modal phonation in English and White Hmong. *Journal of Phonetics 40*:152-161.

Garellek, M. and P. Keating. 2015. Phrase-final creak: Articulation, acoustics and distribution. Conference paper.

Garellek, M., P. Keating, C.M. Esposito and J. Kreiman. 2013. Voice quality and tone identification in White Hmong. *Journal of the Acoustical Society of America 133*:1078-1089.

Garellek, M. and P. Keating. 2011. The acoustic consequences of phonation and tone interactions in Mazatec. *Journal of the International Phonetic Association 41*: 185-205.

Garellek, M. and P. Keating. 2015. Phrase-final creak: Articulation, acoustics, and distribution. Annual Meeting of the Linguistic Society of America, Portland, OR.

Garellek, M., and S. Seyfarth. 2016. Acoustic differences between English /t/ glottalization and phrasal creak. In *Proceedings of Interspeech 2016*. San Francisco. 1054–1058.

Garellek, M., P. Keating, C.M. Esposito and J. Kreiman. 2013. Voice quality and tone identification in White Hmong. *Journal of the Acoustical Society of America 133*: 1078-1089.

Garellek, M., R. A. Samlan, B.R. Gerratt and J. Kreiman. 2016. Modeling the voice source in terms of spectral slopes. *Journal of the Acoustical Society of America 139*: 1404–1410.

Garellek, M., R. A. Samlan, J. Kreiman and B. R. Gerratt. 2013. Perceptual sensitivity to a model of the source spectrum. *Journal of the Acoustical Society of America, Proceedings of Meetings on Acoustics 19*:1-5.

Garellek, M. 2012. The timing and sequencing of coarticulated non-modal phonation in English and White Hmong. *Journal of Phonetics 40*: 152-161.

Garellek, M. 2013. *Production and Perception of Glottal Stops*. Ph.D. thesis: UCLA.

Garellek, M. 2014. Voice quality strengthening and glottalization. *Journal of Phonetics 45*: 106-113.

Garellek, M. 2015. Perception of glottalization and phrase-final creak. *Journal of the Acoustical Society of America 137*: 822-831.

Garellek, M. 2016. The phonetics of voice (draft chapter under consideration for publication in *The Routledge Handbook of Phonetics*). Available online from ResearchGate: https://www.researchgate.net/publication/311558717.

Gerrat, B.R., J. Kreiman and M. Garellek. 2016. Comparing measures of voice quality from sustained phonation and continuous speech. *Journal of Speech, Language, and Hearing Research 59*: 994-1001.

Gerratt, B.R. and J. Kreiman. 2001. Toward a taxonomy of nonmodal phonation. *Journal of Phonetics 29*: 365–381.

Giles, Howard 1979. Ethnicity markers in speech. In Klaus R. Scherer and Giles Howard (eds.) *Social Markers in Speech*. Cambridge, London, New York: Cambridge University Press. 251-289.

Gobl, Christer and Ailbhe Ní Chasaide. 1992. Acoustic characteristics of voice quality. *Speech Communication 11*: 481-490.

Gobl, Christer. 2003. *The Voice Source in Speech Communication: Production and Perception Experiments involving Inverse Filtering and Synthesis*. Stockholm, Sweden. Ph.D. dissertation.

Goodine, Abbey and Alison Johns. 2014. "Would you like fries with thaaaat?" Investigating vocal fry in young female Canadian English speakers. *Strathy Student Working Papers on Canadian English* 2014.

Gordon, Matthew and Peter Ladefoged. 2001. Phonation types: A cross-linguistic overview. *Journal of Phonetics 29*: 383–406.

Gussenhoven, C. 2002. Intonation and interpretation: Phonetics and phonology. In B. Bel and I. Marlien (eds.) *Proceedings of Speech Prosody 2002,* Aix-en Provence, France. 47-57.

Hammarberg, B. 1986. Perceptual and Acoustic Analysis of Dysphonia. Ph.D. thesis. *Studies in Logopedics and Phoniatrics 1,* Stockholm, Sweden: Huddinge University Hospital.

Hanson, Helen M. 1997. Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America 101*: 466–481.

Hanson, Helen M., Kenneth N. Stevens, Hong-Kwang Jeff Kuo, Marilyn Y. Chen and Janet Slifka. 2001. Towards models of phonation. *Journal of Phonetics 29*: 451-480.

Hanson, Helen. M. and Erika S. Chuang. 1999. Glottal characteristics of male speakers: acoustic correlates and comparison with female data. *Journal of the Acoustical Society of America 106*: 1064-1077.

Hawks, John W. and James D. Miller. 1995. A formant bandwidth estimation procedure for vowel synthesis. *Journal of the Acoustical Society of America 97*: 1343-44.

Henderson, E. J. A. 1951. The phonology of loanwords in some South-East Asian languages. *Transactions of the Philological Society*. 131-58.

Henton, Caroline and Anthony Bladon. 1985. Breathiness in a normal female speaker: inefficiency versus desirability. *Language and Communication 5*: 221-7.

Henton, Caroline and Anthony Bladon. 1988. Creak as a sociophonetic marker. In Hyman, L. and Li, C. (eds.) *Language, Speech and Mind*. London: Routledge. 3-29.

Hermes, D.J. 1988. Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America 83*: 257-264.

Hillenbrand, James and Robert A. Houde. 1996. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech, Language and Hearing Research 39*: 311–321.

Hillenbrand, James, Ronald A. Cleveland and Robert L. Erickson. 1994. Acoustic correlates of breathy vocal quality. *Journal of Speech Language and Hearing Research 37*: 769–778.

Hillenbrand, James. 1987. A methodological study of perturbation and additive noise in synthetically generated voice signals. *Journal of Speech and Hearing Research 30*: 448-461.

Hirano, M. 1974. Morphological structure of the vocal cord as a vibrator and its variations. *Folia Phoniatrica (Basel) 26*:89–94.

Hirano, M., J. Ohala and W. Vennard. 1969. The function of laryngeal muscles in regulating fundamental frequency and intensity of phonation. *Journal of Speech and Hearing Research 12*:616–628.

Holliday, Nicole R. and Zachary S. Jaggers. 2015. Influence of suprasegmental features on perceived ethnicity of American politicians. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*. Glasgow.

Hollien, H. and E. Malcik. 1967. Evaluation of cross-sectional studies of adolescent voice change in males. *Speech Monographs 34*:80-84.

Hollien, Harry, Paul Moore, Ronald W. Wendahl and John F. Michel. 1966. On the nature of vocal fry. *Journal of Speech and Hearing Research 9*: 245-247.

Hollien, H., P.A. Hollien and G. DeJong. 1997. Effects of three parameters on speaking fundamental frequency. *Journal of the Acoustical Society of America 102*:2984-2992.

Holmberg, Eva B., Robert E. Hillman and Joseph S. Perkell. 1988. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *Journal of the Acoustical Society of America 84*: 511-529.

Holmberg, Eva B., Robert E. Hillman, Joseph S. Perkell, Peter C. Guiod and Susan L. Goldman. 1995. Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice. *Journal of Speech and Hearing Research 38*:1212 1223.

Honikman, Beatrice. 1964. Articulatory settings. In D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott and J.L.M. Trim (eds.) In *In Honour of Daniel Jones*. London: Longmans. 73-84.

Hudson, A. and A. Holbrook. 1982. Fundamental frequency characteristics of young black adults: Spontaneous speaking and oral reading. *Journal of Speech and Hearing Research 25*:25-28.

Huffman, Marie K. 1987. Measures of phonation type in Hmong. *Journal of the Acoustical Society of America 81*: 495-503.

Hundleby, C.E. 1964. *Xhosa-English Pronunciation in the South-East Cape*. Unpublished Ph.D. thesis: Rhodes University.

Introduction to SAS. UCLA: Statistical Consulting Group. From https://stats.idre.ucla.edu/sas/faq/how-can-i-interpret-log-transformed-variables-in-terms-of-percent-change-in-linear-regression/ (accessed March 10, 2018).

Irwin, R.B. 1977. Judgments of vocal quality, speech fluency, and confidence of southern Black and White speakers. *Language and Speech 20(3)*: 261-266.

Iseli, M. and A. Alwan. 2004. An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation. *Proceedings of ICASSP, vol. 1*, Montreal, Canada. 669–672.

Iseli, M., Y.-L. Shue and A. Alwan. 2006. Age- and gender-dependent analysis of voice source characteristics. *ICASSP*. University of California: Los Angeles. 389-392.

Iseli, M., Y.-L. Shue and A. Alwan. 2007. Age, sex, and vowel dependencies of acoustic measures related to the voice source. *Journal of the Acoustical Society of America 121*: 2283-2295.

Javkin, H., N. Antoñanzas-Barroso and Ian Maddieson. 1987. Digital inverse filtering for linguistic research. *Journal of Speech and Hearing Research 30*: 122-129.

Jessen, Michael and Justus C. Roux. 2002. Voice quality differences associated with stops and clicks in Xhosa. *Journal of Phonetics 30*: 1-52.

Keating, P. and C.M. Esposito. 2006. Linguistic voice quality. *Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology*.

Keating, P., C. M. Esposito, M. Garellek, S. Khan and J. Kuang. 2011. Phonation contrasts across languages. *Proceedings of the International Congress of Phonetic Sciences XVII*, edited by W.-S. Lee and E. Zee. City University of Hong Kong: Hong Kong. 1046-1049.

Keating, P., M. Garellek and J. Kreiman. 2015. Acoustic properties of different kinds of creaky voice. *Proceedings of the 18th International Congress of Phonetic Sciences*.

Kent, Raymond and Charles Read. 2002. *Acoustic Analysis of Speech, 2nd ed*. Australia: Singular/Thomson Learning.

Khan, Sameer Ud Dowla, Kara Becker and Lal Zimman. 2016. The Acoustics of perceived creaky voice in American English. *Journal of the Acoustical Society of America 3*:10.1121. DOI: http://dx.doi.org/10.1121/1.4933741.

Kirk, Paul L., Peter Ladefoged and Jenny Ladefoged. 1984. Using a spectrograph for measures of phonation types in a natural language. *UCLA Working Papers in Phonetics 59*: 102-13.

Kirk, Roger. E. 2013. *Experimental Design: Procedures for the Behavioral Sciences, 4th ed*. Los Angeles, London, New Delhi, Singapore, Washington DC: SAGE Publications.

Klatt, Dennis H and Laura C. Klatt. 1990. Analysis, synthesis, and perception of voice quality variation among female and male talkers. *Journal of the Acoustical Society of America 87*: 820-57.

Knowles, G. 0. 1974. *Scouse, the Urban Dialect of Liverpool*. Unpublished Ph.D. dissertation. Leeds: University of Leeds.

Koike, Y. 1973. Application of some acoustic measures for the evaluation of laryngeal dysfunction. *Studia Phonologica 7*:17-23.

Kreiman, J. and B. Gerratt. 2010. Perceptual sensitivity to first harmonic amplitude in the voice source. *Journal of the Acoustical Society of America 128*:2085-2089.

Kreiman, J. and B. Gerratt. 2011. Modeling overall voice quality with a small set of acoustic parameters. *Journal of the Acoustical Society of America 129*:2529.

Kreiman, J. and B. Gerratt. 2012. Perceptual interaction of the harmonic source and noise in voice. *Journal of the Acoustical Society of America 131*:492-500.

Kreiman, J. and D. Sidtis. 2011. *Foundations of Voice Studies*. Oxford: Wiley-Blackwell.

Kreiman, J., M. Garellek and C.M. Esposito. 2011. Perceptual importance of the voice source spectrum from H2 to 2 kHz. *Journal of the Acoustical Society of America 130*:2570.

Kreiman, J. and B.R. Gerrat. 2006. Predicting perceptual salience. Unpublished manuscript.

Kreiman, J. and B.R. Gerratt. 2010. Perceptual sensitivity to first harmonic amplitude in the voice source. *Journal of the Acoustical Society of America 128*: 2085–2089.

Kreiman, J., B.R. Gerratt and N. Antoñanzas-Barroso. 2007. Measures of the glottal source spectrum. *Journal of Speech, Language, and Hearing Research 50*:595-610.

Kreiman, J., B.R. Gerratt and Sameer ud Dowla Khan. 2010. Effects of native language on perception of voice quality. *Journal of Phonetics 38*: 588–593.

Kreiman, J., B.R. Gerratt, M. Garellek, R. Samlan and Z. Zhang. 2014. Toward a unified theory of voice production and perception. *Loquens*: e009.

Kreiman, J., M. Garellek and C.M. Esposito. 2011. Perceptual importance of the voice source spectrum from H2 to 2 kHz. *Journal of the Acoustical Society of America 130*: 2570.

Krippendorff, Klaus. 2012. Content Analysis: *An Introduction to its Methodology, 3rd ed*. Sage Publications: Thousand Oaks, CA.

Kuang, Jianjing and Mark Liberman. 2015. Influence of spectral cues on the perception of pitch height. In *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK.

Kuang, J. and P. Keating. 2014. Vocal fold vibratory patterns in tense versus lax phonation contrasts. *Journal of the Acoustical Society of America 136*: 2784–2797.

Labov, William. 1963. The social motivation of a sound change. *Word 19*: 273-309.

Labov, William. 1966. *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

Labuschagne, Ilse Bernadette and Valter Ciocca. 2016. The perception of breathiness: Acoustic correlates and the influence of methodological factors. *Journal of Acoustical Science and Technology 37*: 191-201.

Ladefoged, Peter and N. Antoñanzas-Barroso. 1985. Computer measures of breathy phonation. *UCLA Working Papers in Phonetics 61*: 79-86.

Ladefoged, Peter.1971. *Preliminaries to Linguistic Phonetics*. Chicago and London: University of Chicago Press.

Ladefoged, Peter and Ian Maddieson.1996. *The Sounds of the World's Languages*. Malden, Massachusetts: Blackwell Publishers.

Ladefoged, Peter. 1983. The linguistic use of different phonation types. In Diane M. Bless and James H. Abbs (eds.) *Vocal fold physiology: Contemporary Research and Clinical Issues*. San Diego: College-Hill Press. 351-360.

Lass, N.J., C.A. Almerino, L.F. Jordan and J.M. Walsh. 1980. The effect of filtered speech on speaker race and sex identifications. *Journal of Phonetics 8*:101-112.

Lass, N.J., J.E. Tecca, R.A. Mancuso and W.I. Black. 1979. The effect of phonetic complexity on speaker race and sex identifications. *Journal of Phonetics 7*:105-118.

Laver, J., S. Wirz, J. Mackenzie, and S.M. Hiller. 1981. A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress 14*. 139-55.

Laver, John. 1968. Voice quality and indexical information. *British Journal of Disorders of Communication 3*:43-54.

Laver, John. 1972. Voice quality and indexical information. In Laver and Hutcheson (eds.) *Communication in Face to Face Interaction*. Penguin Books. 189-203

Laver, John. 1975. *Individual Features in Voice Quality*. Ph.D. dissertation: University of Edinburgh.

Laver, John. 1980. *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Laver, John. 1991. *The Gift of Speech: Readings in the Analysis of Speech and Voice*. Edinburgh: Edinburgh University Press.

Laver, John and Peter Trudgill. 1979. Phonetic and linguistic markers in speech. In K.R. Scherer and H. Giles (eds.) *Social Markers in Speech*. Cambridge: Cambridge University Press. 1-32.

Lieberman, P. and S. Blumstein. 1988. *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge: Cambridge University Press.

Luchsinger, R. and G.E. Arnold. 1965. Voice speech -language. In *Clinical Communicology : Its Physiology and Pathology*. London: Constable.

Mackenzie Beck, J. M. 1988. *Organic Variation and Voice Quality*. Unpublished Ph.D. dissertation, University of Edinburgh.

Maeda, S. 1993. Acoustics of vowel nasalization and articulatory shifts in French nasalized vowels. In M.K. Huffman and R.A. Krakow (eds.) *Nasals, Nasalization, and the Velum*. San Diego: Academic Press. 147-167.

McCulloch, Charles E. and Shayle R. Searle. 2001. *Generalized, Linear, and Mixed Models*. New York, Chichester, Weinheim, Brisbane, Singapore, Toronto: John Wiley & Sons, Inc.

Mennen, Ineke, James M. Scobbie, Esther de Leeuw, Sonja Schaeffler, and Felix Schaeffler. 2010. Measuring language-specific phonetic settings. *Second Language Research 26(1)*:13-41.

Mesthrie, Rajend. 2010. Socio-phonetics and social change: Deracialisation of the GOOSE vowel in South African English. *Journal of Sociolinguistics 14(1)*: 3-33.

Mesthrie, Rajend. 2016. Race, ethnicity, religion and castes. *The Handbook of Historical Sociolinguistics*.

Mesthrie, Rajend. 2017. Class, gender and substrate erasure in sociolinguistic change: a sociophonetic study of schwa in deracialising South African English. *Language 93(2)*:314-346.

Mesthrie, Rajend, Alida Chevalier and Kate McLachlan. 2015. A perception test for the deracialisation of middle class South African English. *Southern African Linguistics and Applied Language Studies 33(4)*:391-409.

Miller, A.L. 2007. Guttural vowels and guttural coarticulation in Ju|'hoansi. *Journal of Phonetics 35*: 56-84.

Moisik, Scott Reid. 2013. Harsh voice quality and its association with Blackness in popular American media. *Phonetica 69*: 193-215.

Morreira, Kirstin. 2012. *Black South African English: A Sociophonetic Study*. Unpublished Ph.D. dissertation: University of Cape Town.

Morris, R. J. 1997. Speaking fundamental frequency characteristics of 8-through 10-year-old white and African-American boys. *Journal of Communication Disorders 30*:101-116.

Murray, T. 1988. Vocal tract parameters associated with voice quality preference. *Journal of Voice 2*:111-117.

Newman, Michael and Angela Wu. 2011. 'Do you sound Asian when you speak English?' Racial identification and voice in Chinese and Korean Americans' English. *American Speech 86*: 152–178.

Neyman, J. and E.S. Pearson. 1928. On the use and interpretation of certain test criteria for purposes of statistical inference:Part I. *Biometrica 20A(1/2)*: 175-240.

Ng, M. L., Yang Chen and Ellen Y.K. Chan. 2012. Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English Bilingual speakers—A long-term average spectral analysis. *Journal of Voice 26(4)*: e171–e176. doi:10.1016/j.jvoice.2011.07.013

Ng, M., G. Hsueh and C.S. Leung. 2010. Voice pitch characteristics of Cantonese and English produced by Cantonese-English bilingual children. *International Journal of Speech Language Pathology 12*:230-236.

Ní Chasaide, Ailbhe and Christer Gobl.2010. Voice source variation. In William Hardcastle and John Laver (eds.) *The Handbook of Phonetic Sciences*. Cambridge, Massachusetts: Blackwell. 427-461.

Pan, H., M. Chen and S. Lyu. 2011. Electroglottograph and acoustic cues for phonation contrasts in Taiwan Min falling tones. *Proceedings of the 12th INTERSPEECH Conference*, Firenze. 649-652.

Pandit, P.B. 1957. Nasalization, aspiration and murmur in Gujarati. *Indian Linguistics 17*: 165-172.

Perelló, J. 1962. The muco-undulatory theory of phonation. *Annals of Otolaryngology 79*. 722-5

Perrachione, T.K., J.Y. Chiao and P.C.M. Wong. 2010. Asymmetric cultural effects on perceptual expertise underlie and own-race bias for voices. *Cognition 114*:42-55.

Podesva, Robert J. 2007. Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics 11*:478-504.

Podesva, Robert J. 2013. Gender and the social meaning of non-modal phonation types. *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society*. 427-448.

Podesva, Robert J. and Patrick Callier. 2015. Voice quality and identity. *Annual Review of Applied Linguistics 35*: 173-194.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.Rproject.org/.

Rammage, L. A., R.C. Peppard and D.M. Bless. 1992. Aerodynamic, laryngoscopic, and perceptual-acoustic characteristics in dysphonic females with posterior glottal chinks: A retrospective study. *Journal of Voice 6*: 64–78.

Roberts, Michael J. 2008. *Fundamentals of Signals and Systems*. New York: McGraw-Hill.

Rosenfelder, Fruehwald, Evanini and Jiahong 2011. FAVE (Forced Alignment and Vowel Extraction) Program Suite. http://fave.ling.upenn.edu.

Ryan, E. B. and M.A. Carranza. 1977. Ingroup and outgroup reactions to Mexican American language varieties. In Howard Giles (ed.) *Language, Ethnicity and Intergroup Relations*. London: Academic Press.

Samlan, Robin and Brad H. Story. 2011. Relation of structural and vibratory kinematics of the vocal folds to two acoustic measures of breathy voice based on computational modeling. *Journal of Speech, Language, and Hearing Research 54*: 1267-1283.

Samlan, Robin. A., Brad H. Story and Kate Bunton. 2013. Relation of perceived breathiness to laryngeal kinematics and acoustic measures based on computational modeling. *Journal of Speech, Language, and Hearing Research 56*: 1209-1223.

Sapienza, C.M. 1997. Aerodynamic and acoustic characteristics of the adult African voice. *Journal of Voice 11*:410-416.

Sapir, Edward. 1927. Speech as a personality trait. *AJS*. 533-543.

Schielzeth, H. and W. Forstmeier. 2009. Conclusions beyond support: overconfident estimates in mixed models. *Behavioural Ecology 20*:416-420.

Sheskin, David J. 2007. *Handbook of Parametric and Nonparametric Statistical Procedures, 4th ed*. Boca Raton: Chapman & Hall/CRC Taylor & Francis Group.

Shue, Y.-L., P. Keating, C. Vicenik and K. Yu. 2011. Voicesauce: a program for voice analysis. *Proceedings of the ICPhS XVII*. 1846-1849.

Sicoli, M.A. 2007. *Tono: A Linguistic Ethnography of Tone and Voice in a Zapotec region (Mexico).* Unpublished PhD Thesis. Ann Arbor, MI: University of Michigan.

Siegel, Sidney and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences, 2nd ed*. Singapore: McGraw-Hill Book Company.

Simpson, A.P. 2009. Breathiness difference in male and female speech. Is H1-H2 an appropriate measure? *Proceedings of FONETIK*. Stockholm University: Department of Linguistics.

Sjölander, K. 2004. Snack sound toolkit. KTH Stockholm, Sweden. http://www.speech.kth.se/snack.

Slifka, Janet. 2000. *Respiratory Constraints on Speech Production at Prosodic Boundaries*. Ph.D. thesis: MIT.

Slifka, Janet. 2006. Some physiological correlates to regular and irregular phonation at the end of an utterance. *Voice 20*: 171-186.

Smith, Steven W. 2003. *Digital Signal Processing: A Practical Guide for Engineers and Scientists*. Burlington, Massachusetts: Newnes (An imprint of Elsevier).

Södersten, M. and P.Å. Lindestad. 1990. Glottal closure and perceived breathiness during phonation in normally speaking subjects. *Journal of Speech and Hearing Research 33*: 601–611.

Stevens, Kenneth N., G. Fant and S. Hawkins. 1987. Some acoustical and perceptual correlates of nasal vowels. In R. Channon and L. Shockey (eds.) *In Honor of Ilse Lehiste: Ilse Lehiste Püdendusteos*. Dordrecht: Foris. 241-254.

Stevens, Kenneth. N. 1977. Physics of laryngeal behavior and larynx modes. *Phonetica 34*: 264-279.

Stevens, Kenneth N. and Helen M. Hanson. 1995. Classification of glottal vibration from acoustic measurements. In O. Fujimura and M. Hirano (eds.) *Vocal Fold Physiology: Voice Quality Control*. San Diego: Singular. 147-170.

Stevens, Kenneth N. 1998. *Acoustic Phonetics*. Cambridge, Massachusetts: MIT Press.

Stuart-Smith, Jane. 1999. Glasgow: accent and voice quality. In P. Foulkes and G. Docherty (eds.) *Urban Voices: Accent Studies in the British Isles*. London: Arnold. 201-22.

Sun, Xuejing. 2002. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. *Proceedings from ICASSP 2002*. 333-336.

Švec, Jan G., Harm K. Schutte and Donald G. Miller. 1996. A subharmonic vibratory pattern in normal vocal folds. *Journal of Speech, Language and Hearing Research 39*:135-143.

Sweet, Henry. 1877. *Handbook of Phonetics*. Oxford: Clarendon Press.

Sweet, Henry. 1890. *A Primer of Phonetics*. Oxford: Clarendon Press (3rd ed, revised, 1906).

Szakay, Anita 2008. *Ethnic Dialect Identification in New Zealand: The Role of Prosodic Cues*. Saarbrücken, Germany: VDM Verlag Dr Müller.

Szakay, Anita. 2012. Voice quality as a marker of ethnicity in New Zealand: From acoustics to perception. *Journal of Sociolinguistics 16/3*: 382–397.

Szakay, Anita and Eivind Torgersen. 2015. An acoustic analysis of voice quality in London English: The effect of gender, ethnicity and F0. *Proceedings from the 18th International Congress of Phonetic Sciences*, Glasgow.

Tan, Ying Ying. 2012. Age as a factor in ethnic accent identification in Singapore. *Journal of Multilingual and Multicultural Development 33 (6)*: 569–587.

Thomas, E.R. and J. Reaser. 2004. Delimiting perceptual cues used for the ethnic labeling of African American and European American voices. *Journal of Sociolinguistics 8*:54-87.

Titze, I. R. 1988. A framework for the study of vocal registers. *Journal of Voice 2*: 183-194.

Trudgill, Peter. 1974. *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Vetterli, Martin, Jelena Kovačević and Vivek K. Goyal. 2014. *Foundations of Signal Processing*. Cambridge: Cambridge University Press. Vicenik, C., S. Lin, P. Keating and Y.-L., Shue. 2017. Online documentation for VoiceSauce. Available from: http://www.seas.ucla.edu/spapl/voicesauce/index.html

Walton, Julie and Orlikoff, Robert. 1994. Speaker race identification from acoustic cues in the vocal signal. *Journal of American Speech and Hearing Research 37*: 738–745.

Wayland, R., S. Gargash and A. Jongman. 1994. Acoustic and perceptual investigation of breathy voice. *Journal of the Acoustical Society of America 97*:3364.

Wayland, R. and A. Jongman. 2003. Acoustic correlates of breathy and clear vowels: the case of Khmer. *Journal of Phonetics 31*: 181-201.

Wells, John. 1982. *Accents of English, Vol. 1–3*. Cambridge: Cambridge University Press.

Winter, B. 2013. Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. [http://arxiv.org/pdf/1308.5499.pdf].

Wissing, D. 2002. Black South African English: A new English? Some observations from a phonetic viewpoint. *World Englishes 21*:129-144.

Xue, S.A. and D. Fucci. 2000. Effects of sex and race on acoustic features of voice analysis. *Perceptual and Motor Skills 91*:951-958.

Xue, S.A. and J. G. Hao. 2006. Normative standards for vocal tract dimension by race as measured by acoustic pharyngometry. *Journal of Voice 20*:391-400.

Xue, S.A., G.J. Hao and R. Mayo. 2006. Volumetric measurements of vocal tracts for male speakers from different races. *Clinical Linguistics and Phonetics 20*:691-702.

Xue, S.A., R. Needley, F. Hagstrom and J. Hao. 2001. Speaking F0 characteristics of elderly Euro-American and African-American speakers: building a clinical comparative platform. *Journal of Clinical Linguistics and Phonetics 15*:245-252.

Yamamura, Kohji. 1999. Transformation using (x + 0.5) to stabilize the variance of populations. *Researches on Population Ecology 41*: 229-234.

Yuasa, Ikuko Patricia. 2010. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech 85*: 315-337.

Yumoto, E. 1983.The quantitative evaluation of hoarseness: A new harmonics to noise ratio method. *Archives of Otolaryngology 109*: 48-52.

Yumoto, E., W.J. Gould and T. Baer. 1982. Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America 71*: 1544-1550.

Yumoto, E., Y. Sasaki and H. Okamura, H. 1984. Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. *Journal of Speech and Hearing Research 27*:2-6.

Zhang, Z., J. Kreiman and B.R. Gerratt. 2011. Perceptual sensitivity to changes in vocal fold geometry and stiffness. *Journal of the Acoustical Society of America 129*: 2529.

Zhang, Z., J. Kreiman, B.R. Gerratt and M. Garellek. 2013. Acoustic and perceptual effects of changes in body layer stiffness in symmetric and asymmetric vocal fold models. *Journal of the Acoustical Society of America 133*: 453–462.

Zhang, Z. 2015. Regulation of glottal closure and airflow in a three-dimensional phonation model: Implications for vocal intensity control. *Journal of the Acoustical Society of America 137*: 898–910.

Zhang, Z. 2016a. Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model. *Journal of the Acoustical Society of America 139*: 1493–1507.

Zhang, Z. 2016b. Mechanics of human voice production and control. *Journal of the Acoustical Society of America 140*: 2614–2635.

Zimman, Lal. 2013. Hegemonic masculinity and the variability of gay-sounding speech: The perceived sexuality of transgender men. *Journal of Language and Sexuality 2(1)*: 1–39. doi:10.1075/jls.2.1.01zim .