COMPUTERISED GRBAS ASSESSEMENT OF VOICE QUALITY

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2016

By
Farideh Jalalinajafabadi
School of Computer Science

Contents

Abstract **22**

Declaration **24**

Copyright **25**

Acknowledgments **26**

1 Introduction **28**

1.1 Research Hypothesis 30

1.2 Aim and Objectives 30

1.3 Research Contributions 31

1.4 Publications 32

1.5 Thesis Structure 34

2 Background and Related Work **36**

2.1 Introduction 36

2.2 Speech and human speech production 37

2.2.1 Vocal folds 39

2.2.2 Vocal tract 41

2.2.3 Speech phonemes 42

2.3 Voice disorder 43

2.4 Voice quality analysis 45

2.4.1 Voice Handicap Index 46

2.4.2 GRBAS 46

2.4.3 Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) 48

2.5 Objective voice measurement 50

2.6 ‘Praat’ software package 54

2.6.1 'Praat' measurement of fundamental frequency 54
2.6.2 'Praat' measurements of Jitter 54
2.6.3 Praat measurements of noise-based features 58
2.7 Multi-Dimensional Voice Program (MDVP) 58
2.7.1 'MDVP' measurements of fundamental frequency 59
2.7.2 'MDVP' measurements of jitter 59
2.7.3 'MDVP' measurements of shimmer 60
2.7.4 'MDVP' measurements of noise-based features 61
2.8 Analysis of Dysphonia in Speech and Voice (ADSV) 63
2.9 Discussion 65
2.10 Other software packages for speech analysis 65
2.11 Comparisons between Praat, MDVP and other packages 65
2.12 Use of existing software packages in this thesis 68
2.13 Machine learning algorithms 69
2.13.1 Type of learning 69
2.13.2 How supervised learning algorithms work 70
2.13.3 Cross-Validation 71
2.14 Objective voice quality analysis conforming to the GRBAS scale . . 72
2.15 Conclusions 74

3 Data-base Creation and GRBAS Scoring 76

3.1 Introduction 76
3.2 Data-base creation 77
3.3 GRBAS scoring 78
3.3.1 Introduction and motivation 78
3.3.2 Automated GRBAS scoring 79
3.4 Reliability testing of GRBAS scoring 81
3.4.1 Pearson correlation 82
3.4.2 Root Mean Square Error 83
3.4.3 Intra-Class Correlation (ICC) 84
3.5 Kappa measurements 85
3.5.1 Cohen's Kappa 86
3.5.2 Weighted Kappa 88
3.5.3 Fleiss' Kappa 90
3.5.4 Fleiss Kappa demystified 92
3.5.5 Weighted Fleiss Kapppa and Farideh's Kappa 94

3.5.6 Intra-Scorer Consistency 96
3.6 Conclusions so far 99
3.6.1 Inter-Scorer Consistency 100
3.7 Reference GRBAS scores 102
3.7.1 Averages 102
3.8 'Gold standard' GRBAS reference scores 105
3.9 Conclusions 107

4 Feature Measurement 109
4.1 Introduction 109
4.2 Some Background 110
4.2.1 Recording and digitising Speech 110
4.2.2 Voiced and unvoiced speech 111
4.2.3 Analysing voiced speech 111
4.2.4 Sound due to turbulent air-flow 112
4.3 DSP techniques for measuring voice features 114
4.3.1 Fast Fourier Transform (FFT) methods 114
4.3.2 Short term Fourier transform (STFT) 115
4.3.3 Cepstrum methods 118
4.3.4 Autocorrelation based methods 124
4.3.5 Cross-correlation methods 126
4.3.6 Measuring degrees of periodicity using the cross-correlation . 128
4.3.7 Aperiodicity Index (API) and harmonic to noise ratio (HNR) . 132
4.3.8 Enhanced cross-correlation method 132
4.3.9 Advantages of the cross-correlation method 133
4.4 Measuring fundamental frequency, voicing, jitter and shimmer 134
4.5 Measuring jitter for artificial voiced speech 135
4.6 Measuring shimmer for artificial voiced speech 137
4.7 Measuring artificial voiced speech with jitter and shimmer 138
4.8 Measuring Harmonic-to-Noise Ratio (HNR) 140
4.9 Measuring HNR for artificial voiced speech 145
4.10 Measuring other features 146
4.11 Feature measurement by ADSV 148
4.12 Identifying features likely to be indicative of GRBAS components. . . 151
4.12.1 Grade 151
4.12.2 Roughness 151

4.12.3 Breathiness 151
4.12.4 Asthenia 152
4.12.5 Strain 152
4.13 Conclusions 152

5 Objective Prediction of GRBAS Scores 154
5.1 Supervised learning prediction models 155
5.1.1 'Multiple Linear Regression' (MLR) 155
5.1.2 'K Nearest Neighbour' (KNN) Regression 157
5.1.3 Performance of prediction models 160
5.2 Training and testing 161
5.2.1 Optimum K for the GRBAS prediction by KNNR 162
5.3 Prediction by all features 165
5.3.1 Discussion 166
5.4 Progressively discarding features 167
5.4.1 Improvement in performance of the prediction models 173
5.5 Dimensionality reduction 173
5.5.1 Filter methods 174
5.5.2 Feature ranking 175
5.5.3 Feature reduction 175
5.5.4 Wrapper methods 175
5.6 Grade Prediction 179
5.6.1 Greedy forward selection for 'Grade' 179
5.6.2 Exhaustive search for 'Grade' 181
5.6.3 Optimum K for Grade prediction by KNNR 185
5.7 Roughness prediction 186
5.7.1 Greedy forward selection for 'Roughness' 186
5.7.2 Exhaustive search for 'Roughness' 186
5.7.3 Optimum K for Roughness prediction by KNNR 189
5.8 Breathiness prediction 189
5.8.1 Greedy forward selection for 'Breathiness' 189
5.8.2 Exhaustive search for 'Breathiness' 191
5.8.3 Optimum K for Breathiness prediction by KNNR 191
5.9 Asthenia Prediction 193
5.9.1 Greedy forward selection for 'Asthenia' 193
5.9.2 Exhaustive search for 'Asthenia' 195

5.9.3 Optimum K for Asthenia prediction by KNNR 197
5.10 Strain Prediction 198
5.10.1 Greedy forward selection for ‘Strain’ 198
5.10.2 Exhaustive search for ‘Strain’ 199
5.10.3 Optimum K for Strain prediction by KNNR 201
5.11 Exhaustive search compared with greedy forward selection 202
5.12 Comparison between MLR and KNNR 202
5.13 Conclusions 203

6 Results and Evaluation 206
6.1 Introduction 206
6.2 Objective scoring compared with SLT scores 207
6.2.1 MLR and KNNR for ‘Grade’ scoring 207
6.2.2 MLR and KNNR for ‘Roughness’ scoring 208
6.2.3 MLR and KNNR for ‘Breathiness’ scoring 209
6.2.4 MLR and KNNR for ‘Asthenia’ scoring 209
6.2.5 MLR and KNN for ‘Strain’ scoring 210
6.3 Conclusions 211

7 Conclusions and Suggestions for Further Work 213
7.1 Suggestions for further Work 217

Bibliography 220

A Ethical approval 235

B Data-base Definition 244

C ‘Gold-standard’ Reference Scores 245

D Measuring Fundamental Frequency (FF) for artificial voiced speech 248

E Definitions of Feature Subsets used in Chapter 5 250

List of Tables

3.1 Significance of Pearson Correlation Coeffs 82

3.2 Illustration of possible scores for Scorer 1 and Scorer 2 83

3.3 Significance of ICC values 85

3.4 Score-distribution table 87

3.5 Significance of Kappa values 88

3.6 Examples of Fleiss Kappa. 93

3.7 The example of original and new linear weightings given by the new Fleiss Kappa. 95

3.8 Mean of GRBAS scores for 20 random examples scored twice 97

3.9 Intra-Scorer Consistency in Grade 97

3.10 Intra-Scorer Consistency in Roughness 98

3.11 Intra-Scorer Consistency in Breathiness 98

3.12 Intra-Scorer Consistency in Asthenia 98

3.13 Intra-Scorer Consistency in Strain 98

3.14 Aggregate of measurement of consistency for each GRBAS component over all scorers 99

3.15 Aggregate of measurement of consistency for each scorer over all 5 GRBAS components 100

3.16 Means of GRBAS Scores 100

3.17 Inter-Scorer Agreement in GRBAS scoring 101

3.18 Kappa & weighted Kappa averaged over all SLT pairs in GRBAS scoring101

3.19 Fleiss Kappa linearly weighted Fleiss Kappa for 5 GRBAS 102

3.20 ICC between the five SLTs for GRBAS 102

4.1 Comparison of RL Jitter measurements for artificial voiced speech . . 136

4.2 Comparison of Shimmer measurements for artificial voiced speech . . 139

4.3 Jitter & Shimmer for artificial voiced speech with SNR = infinity . . . 141

4.4 Jitter & Shimmer for artificial voiced speech with SNR = 10dB 142

4.5 Comparison of HNR measurements for periodic waveform (4.18) with added noise 144

4.6 Comparison of HNR measurements for artificial voiced speech with added noise 146

4.7 Features measured by thesis software 148

4.8 Features measured by ADSV. 149

5.1 Illustration of five nearest neighbours 158

5.2 Prediction results using all 20 features. 166

5.3 Prediction Results for Grade by MLR and KNNR 168

5.4 Prediction Results for Roughness by MLR and KNNR 168

5.5 Prediction Results for Breathiness by MLR and KNNR 169

5.6 Prediction Results for Asthenia by MLR and KNNR 169

5.7 Prediction Results for Strain by MLR and KNNR 170

5.8 Correlation of individual feature measurements with Grade score. . . 170

5.9 Correlation of individual feature measurements with Roughness score. 171

5.10 Correlation of individual feature measurements with Breathiness score. 171

5.11 Correlation of individual feature measurements with Asthenia score. . 172

5.12 Correlation of individual feature measurements with Strain score. . . . 172

5.13 Comparison Between MLR And KNNR for All twenty Features . . . 203

5.14 Comparison Between MLR And KNNR for the Best Subset 203

C.1 ‘Gold-standard’ ref scores compared with unweighted averages (PART 1). 245

C.2 ‘Gold-standard’ ref scores compared with unweighted averages (PART 2). 246

C.3 ‘Gold-standard’ ref scores compared with unweighted averages (PART 3). 247

D.1 Comparison of Fundamental Frequency (FF) measurements for artificial voiced speech 248

E.1 Forward Selection subsets referred to in Figures 5.8 and 5.9 (Grade) . 251

E.2 Exhaustive search subsets referred to in Figures 5.10 and 5.11 (Grade) 252

E.3 Forward Selection subsets referred to in Figures 5.15 and 5.16 (Roughness) 253

E.4 Exhaustive search subsets referred to in Figures 5.18 and 5.17 (Roughness) . . . 254
E.5 Forward Selection subsets referred to in Figures 5.22 and 5.23 (Breathiness) . . . 255
E.6 Exhaustive search subsets referred to in Figures 5.24,5.25 (Breathiness) 256
E.7 Forward Selection subsets referred to in Forward Selection subsets referred to in Figures 5.29 and 5.30 (Asthenia) . . . 257
E.8 Exhaustive search subsets referred to in Figures 5.31 and 5.32 (Asthenia)258
E.9 Forward Selection subsets referred to in Figures 5.36 and 5.37 (Strain) 259
E.10 Exhaustive search subsets referred to in Figures 5.38 and 5.39 (Strain) 260

List of Figures

1.1 the goal of the project. 30

2.1 Speech Production Model 38

2.2 EGG device applied to a patient's larynx 40

2.3 EGG waveform produced by short 45.35 ms segment of the sustained vowel /a/. The sampling frequency is 44100 Hz 41

2.4 Speech waveform produced by short 45.35 ms segment of the sustained vowel /a/. The sampling frequency is 44100 Hz 41

2.5 VHI-10 voice is a self-assessment of voice quality by participants who attend voice clinics. The VHI with 30 items is evaluated with the range of responses: 'never', 'almost never', 'sometimes', 'almost always', and 'always'. In analysing the questionnaire, each response was scored from 0 to 4. The, total score could range from 0 to 120 points. 47

2.6 A graphical user interface (GUI) in supplied by Kay-Pentax [Kay96] represents CAPE-V attributes (Overall severity, Roughness, Breathiness, Strain) by a 100-millimeter line forming a visual analog scale (VAS).'MI' refers to 'mildly deviant', 'MO' refers to 'moderately deviant' and 'SE' refers to 'severely deviant'. 50

2.7 This is the CAPE-V form. Clinicians measures overall severity, roughness, breathiness, strain, pitch, and loudness out of 100.'MI' refers to 'mildly deviant', 'MO' refers to 'moderately deviant' and 'SE' refers to 'severely deviant'. 51

2.8 This is a windowing containing ADSV data. The green data mark at the start of the data. The red data cursor has been moved to the instance of L/H Ratio and CPP data located. This is reported in the x-axis box located on the status line. The waveform, L/H Ratio contour, and CPP contour are displayed. ADSV Results are calculated from the blue highlighted data located between the selection cursors. 64

3.1 This is a Screen shot of the GPSP. SLT scores the GRBAS attributes from 0 to 3. The scorer name in is provided in the text box. Each SLT can observe firstly, the total number recordings that should be scored, secondly, the recordings that have been scored and thirdly the remaining recordings that have not been scored. Five buttons for ‘open Datbase’, Listen/Next, Listen again, submits scores, and save and exist are provided. 80

3.2 Histogram of differences between gold-standard and unweighted arithmetic mean for Grade as listed in Appendix B. Both rounded to one decimal place. The frequency is the number of subject out of the total of 102 105

3.3 Histogram of differences between gold-standard and unweighted arithmetic mean for Roughness as listed in Appendix B. Both rounded to one decimal place. The frequency is the number of subject out of the total of 102 106

3.4 Histogram of differences between gold-standard and unweighted arithmetic mean for Breathiness as listed in Appendix B. Both rounded to one decimal place. The frequency is the number of subject out of the total of 102 106

3.5 Histogram of Differences between gold-standard and unweighted arithmetic mean for Asthenia as listed in Appendix B. Both rounded to one decimal place. The frequency is the number of subject out of the total of 102 107

3.6 Histogram of Differences between gold-standard and unweighted arithmetic mean for Strain as listed in Appendix B both rounded to one decimal place. The frequency is the number of subject out of the total of 102 107

4.1 Synthetic vowel in time domain (a) and frequency domain (b). The time-domain sample index refers to the speech sampled at 8 kHz. The FFT frequency index refers to the 512 point FFT of the synthetic vowel. The frequency represented by k is $F_s \times k/512$ where F_s= 8kHz. The red graph is the impulse-response. 119

4.2 Real cepstrum for artificial speech. The Cepstrum index k represents time in units of 512/ F_s= 64 ms. Therefore k represents (64k) ms in time (or quefrequency). The red graph is the impulse-response. . . . 120

4.3 Synthetic vowel with noise time domain (a) and cepstrum (b). The time-domain sample index refers to the speech sampled at 8 kHz. The Cepstrum index k represents time in units of 512/ F_s= 64 ms. Therefore k represents (64k) ms in time (or quefrequency). The red graph is the impulse-response. 121

4.4 Synthetic vowel with noise and jitter Time domain (a) and Cepstrum (b). The time-domain sample index refers to the speech sampled at 8 kHz. The Cepstrum index k represents time in units of 512/ F_s= 64 ms. Therefore k represents (64k) ms in time (or quefrequency). The red graph is the impulse-response 123

4.5 Synthetic vowel with amplitude modulation. The time-domain sample index refers to the speech sampled at 8 kHz. The red graph is the impulse-response. 124

4.6 Synthetic vowel with amplitude and noise time domain (a) and cepstrum (b). The time-domain sample index refers to the speech sampled at 8 kHz. The Cepstrum index k represents time in units of 512/ F_s= 64 ms. Therefore k represents (64k) ms in time (or quefrequency). The red graph is the impulse-response. 125

4.7 HNR measurements for a periodic signal of fundamental frequency 200 Hz sampled at 40 kHz waveform with added noise compared with SNR 130

4.8 Comparison of RL Jitter measurements for artificial voiced speech for thesis software and Praat software 137

4.9 Comparison of RL Shimmer measurements for artificial voiced speech for thesis software and Praat software 138

4.10 thesis-HNR & Praat-HNR compared with SNR for periodic waveform for thesis software and Praat software 145

4.11 Comparison of HNR measurements for artificial voiced speech with added noise for thesis software, Praat and MDVP 147

5.1 Full cross-validation procedure, with train, validation and test sets. . . 162

5.2 Average of Trial-Averaged-Validation-NRMSE for K=1 to K=10 (Grade) 163

5.3 Average of Trial-Averaged-Validation-NRMSE for K=1 to K=10 (RBAS) 164

5.4 Optimum K for 20 features in each Trial (Grade) 164

5.5 Optimum K for 20 features in each Trial (RBAS). 165

5.6 Greedy forward selection by 10-fold cross validation. The procedure started with an empty subset and incrementally introduced features one by one. 177

5.7 Feature Selection by exhaustive search. In each trial 10-Fold cross validation was carried out for 1023 different subsets of features. . . . 179

5.8 Trial-Testing-NRMSE for the best selected feature subset (MLR-Grade-Greedy). S labels are features subsets in Table E.1 180

5.9 Trial-Testing-NRMSE for the best selected feature subset (KNNR-Grade-Greedy). S labels are features subsets in Table E.1 181

5.10 Average of Trial-Testing-NRMSE for the best 20 selected feature (MLR-Grade-EXHAUSTIVE). S labels are feature subsets in Table E.2. (‘o’:average, ‘-’:median, ‘+’:outliers) 182

5.11 Average of Trial-Testing-NRMSE for the best 20 selected feature subset (KNNR-Grade-EXHAUSTIVE). S labels are feature subsets in Table E.2. (‘o’:average, ‘-’:median, ‘+’:outliers) 183

5.12 Average of Trial-Testing-NRMSE for all subset of features (MLR-Grade-EXHAUSTIVE). (green circle : best subset and red circle: 10 features) 184

5.13 Average of Trial-Testing-NRMSE for all subset of features (KNNR-Grade-EXHAUSTIVE). (green circle : best subset and red circle: 10 features) 184

5.14 Average of Trial-Averaged-Val-NRMSE Per K Best Subset (Grade) . 185

5.15 Trial-Testing-NRMSE for the best selected feature subset (MLR-Roughness-Greedy). S labels are features subsets in Table E.3 186

5.16 Trial-Testing-NRMSE for the best selected feature subset (KNNR-Roughness-Greedy). S labels are features subsets in Table E.3 187

5.17 Average of Trial-Testing-NRMSE for the best 20 selected feature subset (MLR-Roughness-EXHAUSTIVE). S labels are feature subsets in Table E.4. (‘o’:average, ‘-’:median, ‘+’:outliers) 187

5.18 Average of Trial-Testing-NRMSE for the best 20 selected feature subset (KNNR-Roughness-EXHAUSTIVE). S labels are feature subsets in Table E.4. (‘o’:average, ‘-’:median, ‘+’:outliers) 188

5.19 Average of Trial-Testing-NRMSE for all subset of features (MLR-Roughness-EXHAUSTIVE). (green circle: best subset and red circle: 10 features) 188

5.20 Average of Trial-Testing-NRMSE for all subset of features (KNNR-Roughness-Exhaustive). (green circle: best subset and red circle: 10 features) 189

5.21 Average of Trial-Averaged-Val-NRMSE Per K for Best Subset (Roughness) 190

5.22 Trial-Testing_NRMSE for the best selected feature subset (MLR-Breathiness-Greedy). S labels are features subsets in Table E.5 190

5.23 Trial-Testing_NRMSE for the best selected feature subset (KNNR-Breathiness-Greedy). S labels are features subsets in Table E.5 191

5.24 Average of Trial-Testing-NRMSE for the best 20 selected feature subset in each trial (MLR-Breathniess-EXHAUSTIVE). S labels are feature subsets in Table E.6. (‘o’:average, ‘-’:median, ‘+’:outliers) 192

5.25 Average of Trial-Testing-NRMSE for the best 20 selected feature subset in each trial (KNNR-Breathniess-EXHAUSTIVE). S labels are feature subsets inTable E.6. (‘o’:average, ‘-’:median, ‘+’:outliers) 192

5.26 Average of Trial-Testing-NRMSE for all subset of features (MLR-Breathiness-EXHAUSTIVE). (green circle : best subset and red circle: 10 features) 193

5.27 Average of Trial-Testing-NRMSE for all subset of features (KNNR-Breathiness-EXHAUSTIVE). (green circle : best subset and red circle: 10 features) 193

5.28 Average of Trial-Averaged-Val-NRMSE Per K for Best Subset (Breathiness) 194

5.29 Trial-Testing-NRMSE for the best selected feature subset (MLR-Asthenia-Greedy). S labels are features subsets in Table E.7 194

5.30 Trial-Testing-NRMSE for the best selected feature subset (KNNR-Asthenia-Greedy). S labels are features subsets in Table E.7 195

5.31 Average of Trial-Tesing-NRMSE for the best 20 selected feature subset (MLR-Asthenia-EXHAUSTIVE). S labels are feature subsets in Table E.8. (‘o’:average, ‘-’:median, ‘+’:outliers) 196

5.32 Average of Trial-Tesing-NRMSE for the best 20 selected feature subset (KNNR-Asthenia-EXHAUSTIVE). S labels are feature subsets in Table E.8. (‘o’:average, ‘-’:median, ‘+’:outliers) 196

5.33 Average-Trial-Testing-NRMSE for all subset of features (MLR-Asthenia-EXHAUSTIVE). (green circle: best subset and red circle: 10 features) 197

5.34 Average of Trial-Testing-NRMSE for all subset of features (KNNR-Asthenia-EXHAUSTIVE). (green circle : best subset and red circle: 10 features) 197

5.35 Average of Trial-Averaged-Val-NRMSE Per K for Best Subset (Asthenia) 198

5.36 Trial-Testing-NRMSE for the best selected feature subset (MLR-Strain-Greedy). S labels are features subsets in Table E.9 198

5.37 Trial-Testing-NRMSE for the best selected feature subset (KNNR-Strain-Greedy). S labels are features subsets in Table E.9 199

5.38 Average of Trial-Testing-NRMSE for the best 20 selected feature subset (MLR-Strain-EXHAUSTIVE). S labels are feature subsets in Table E.10. ('o':average, '-':median, '+':outliers) 200

5.39 Average of Trial-Testing-NRMSE for the best 20 selected feature subset (KNNR-Strain-EXHAUSTIVE). S labels are feature subsets in Table E.10. ('o':average, '-':median, '+':outliers) 200

5.40 Average of Trial-Testing-NRMSE for all subset of features. green circle best subset (KNNR-Strain-EXHAUSTIVE). (green circle: best subset and red circle: 10 features) 201

5.41 Average of Trial-Testing-NRMSE for all subset of features (KNNR-Strain-EXHAUSTIVE). (green circle: best subset and red circle: 10 features) 201

5.42 Average of Trial-Averaged-Val-NRMSE Per K for Best Subset (Strain) 202

6.1 Average of Trial-Testing-NRMSE for subjective and objective 'Grade' scoring. The observed Grade scores of individual SLT and the predicted Grade scores of objective systems (MLR and KNNR) have been compared against the average of scores of five scorers for the 22 'set-aside' testing subjects in each trial 208

6.2 Average of Trial-Testing-NRMSE for subjective and objective 'Roughness' scoring. The observed 'Roughness' scores of individual SLT and the predicted Roughness scores of objective systems (MLR and KNNR) have been compared against the average of scores of five scorers for the 22 'set-aside' testing subjects in each trial 209

6.3 Average of Trial-Testing-NRMSE for subjective and objective 'Breathiness' scoring. The observed 'Breathiness' scores of individual SLT and the predicted Breathiness scores of objective systems (MLR and KNNR) have been compared against the average of scores of five scorers for the 22 'set-aside' testing subjects in each trial 210

6.4 Average of Trial-Testing-NRMSE for subjective and objective 'Asthenia' scoring. The observed 'Asthenia' scores of individual SLT and the predicted Asthenia scores of objective systems (MLR and KNNR) have been compared against the average of scores of five scorers for the 22 'set-aside' testing subjects in each trial 210

6.5 Average of Trial-Testing-NRMSE for subjective and objective 'Strain' scoring. The observed 'Strain' scores of individual SLT and the predicted Strain scores of objective systems (MLR and KNNR) have been compared against the average of scores of five scorers for the 22 'set-aside' testing subjects in each trial 211

Acronyms

ADSV Analysis of Dysphonia in Speech and Voice. 27

ANN Artificial Neural Network. 71

API Aperiodicity Index. 130

APQ3 Three-Point Amplitude Perturbation Quotient. 55

APQ5 Five-Point Amplitude Perturbation Quotient. 55

CAPE-V Consensus Auditory-Perceptual Evaluation of Voice. 46

CPP Cepstral Peak Prominence. 111

CSID Cepstral/Spectral Index of Dysphonia. 107

DFT Discrete Fourier Transform. 112

DSP Digital Signal Processing. 20

EGG Electro-glottograph. 38

FFT Fast Fourier Transform. 59

GENR Glottal Excitation to Noise Ratio. 51

GMM Gaussian Mixture Model. 71

GPSP GRBAS Presentation and Scoring Package. 21

HMM Hidden Markov Model. 71

HNR Harmonic to Noise Ratio. 51

ICC Intra-class correlation. 79

KNNR K Nearest Neighbor Regression. 21

LVQ Learning Vector Quantization. 70

MDVP Multi-Dimensional-Voice-Program. 27

MFCCs Mel Frequency Cepstral Coefficients. 43

MLP Multi-Layer Perceptron. 71

MLR Multiple Linear Regression. 21

MRI Manchester Royal Infirmary. 20

NNE Normalised Noise Energy. 51

NRMSE Normalised Root Mean Squared Error. 21

PCA Principal Components Analysis. 140

PCR Principal Component Regression. 165

PLSR Partial Least Squares Regression. 165

PPQ5 Five-Point Period Perturbation Quotient. 54

PPV Pitch- Period Variation. 133

RAP Relative Average Perturbation. 53

RL Relative-Local. 133

RMSE Root Mean Square Error. 81

SAPQ Smoothed Amplitude Perturbation Quotient. 58

SHV Shimmer Variation. 135

SLTs Speech and Language Therapists. 20

SNR Signal to Noise Ratio. 128

SPPQ Smoothed Pitch Period Perturbation Quotient. 57

STFT Short-Time Fourier Transform. 113

SVM Support Vector Machine. 71

VHI Voice Handicap Index. 44

VoIP Voice over IP. 110

VTI Voice Turbulence Index. 59

Nomenclature

$1-C_{\mathrm{max}}$	The degree of non-periodicity
$\bar{P}$	The average of $\mathrm{P_i}$
β	The regression coefficients
σ	The standard deviation
A_{i}	The consecutive pitch-cycles
C	Th scoring categories
D_{e}	The probability of disagreement by chance
D_{ij}	The number of scorers pair that differ
D_{o}	The proportion of subject for which the two scorers disagree
E	The sum of squares
K	The number of the neighbors in KNNR
p_{e}	The probability of agreement by chance
P_{i}	The proportion of pairs of scorers that agree for that subject
p_{o}	The proportion of subject for which the two scorers agree
R_{M}	The Reliability Matrix
s_{f}	The smoothing factor
T_{i}	The pitch period
W_{L}	The linear weighting matrix

W_{m}	The weighted arithmetic means
W_{S}	The non-Linear squared matrix
W_{U}	The linear unweighting matrix
X_{k}	The kth frequency domain coefficient
$X^{\#}$	The pseudo inverse

Abstract

COMPUTERISED GRBAS ASSESSEMENT OF VOICE QUALITY
Farideh Jalalinajafabadi
A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy, 2016

Vocal cord vibration is the source of voiced phonemes in speech. Voice quality depends on the nature of this vibration. Vocal cords can be damaged by infection, neck or chest injury, tumours and more serious diseases such as laryngeal cancer. This kind of physical damage can cause loss of voice quality. To support the diagnosis of such conditions and also to monitor the effect of any treatment, voice quality assessment is required. Traditionally, this is done 'subjectively' by Speech and Language Therapists (SLTs) who, in Europe, use a well-known assessment approach called 'GRBAS'.

GRBAS is an acronym for a five dimensional scale of measurements of voice properties. The scale was originally devised and recommended by the Japanese Society of Logopeadics and Phoniatrics and several European research publications. The properties are 'Grade', 'Roughness', 'Breathiness', 'Asthenia' and 'Strain'. An SLT listens to and assesses a person's voice while the person performs specific vocal maneuvers. The SLT is then required to record a discrete score for the voice quality in range of 0 to 3 for each GRBAS component. In requiring the services of trained SLTs, this subjective assessment makes the traditional GRBAS procedure expensive and time-consuming to administer.

This thesis considers the possibility of using computer programs to perform objective assessments of voice quality conforming to the GRBAS scale. To do this, Digital Signal Processing (DSP) algorithms are required for measuring voice features that may indicate voice abnormality. The computer must be trained to convert DSP measurements to GRBAS scores and a 'machine learning' approach has been adopted to achieve this. This research was made possible by the development, by Manchester Royal Infirmary (MRI) Hospital Trust, of a 'speech database' with the participation of clinicians, SLT's, patients and controls. The participation of five SLTs scorers allowed norms to be established for GRBAS scoring which provided 'reference' data for the

machine learning approach.

To support the scoring procedure carried out at MRI, a software package, referred to as GRBAS Presentation and Scoring Package (GPSP), was developed for presenting voice recordings to each of the SLTs and recording their GRBAS scores. A means of assessing intra-scorer consistency was devised and built into this system. Also, the assessment of inter-scorer consistency was advanced by the invention of a new form of the 'Fleiss Kappa' which is applicable to ordinal as well as categorical scoring. The means of taking these assessments of scorer consistency into account when producing 'reference' GRBAS scores are presented in this thesis. Such reference scores are required for training the machine learning algorithms.

The DSP algorithms required for feature measurements are generally well known and available as published or commercial software packages. However, an appraisal of these algorithms and the development of some DSP 'thesis software' was found to be necessary. Two 'machine learning' regression models have been developed for mapping the measured voice features to GRBAS scores. These are K Nearest Neighbor Regression (KNNR) and Multiple Linear Regression (MLR). Our research is based on sets of features, sets of data and prediction models that are different from the approaches in the current literature.

The performance of the computerised system is evaluated against reference scores using a Normalised Root Mean Squared Error (NRMSE) measure. The performances of MLR and KNNR for objective prediction of GRBAS scores are compared and analysed 'with feature selection' and 'without feature selection'. It was found that MLR with feature selection was better than MLR without feature selection and KNNR with and without feature selection, for all five GRBAS components.

It was also found that MLR with feature selection gives scores for 'Asthenia' and 'Strain' which are closer to the reference scores than the scores given by all five individual SLT scorers. The best objective score for 'Roughness' was closer than the scores given by two SLTs, roughly equal to the score of one SLT and worse than the other two SLT scores. The best objective scores for 'Breathiness' and 'Grade' were further from the reference scores than the scores produced by all five SLT scorers. However, the worst 'MLR with feature selection' result has normalised RMS error which is only about 3% worse than the worst SLT scoring.

The results obtained indicate that objective GRBAS measurements have the potential for further development towards a commercial product that may at least be useful in augmenting the subjective assessments of SLT scorers.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgments

In the name of ‘God’ who gave me motivation, courage during this research work.

I would like to thank to my supervisors, Dr. Barry Cheetham and Dr. Mikel Luján for their support during these four years. More specifically, my biggest thank goes to Barry who expertly guided me through of these four years. Barry was a fantastic supervisor, he understood my strengths and weaknesses from early on, and the implausible faith he showed in me is what has kept me motivated throughout these years.

I would like to say thanks to my collaborators in Manchester Royal Infirmary Hospital Prof. Jarrod Homer, Dr. Chaitanya Gadepalli and Frances Ascott. More specially, Dr. Chaitanya Gadepalli for his guidance and support throughout my research. He contributed enormously to parts of Chapter 3.

My thanks is also expressed to the patients and volunteers in MRI who made a unique data base possible for this research.

Thanks to Dr. Simon Harper, Dr. Gavin Brown, Dr. Eva Navarro, and Dr. Konstantinos Sechidis for their kind support. I feel blessed for having met them on my path.

I am thankful to my colleagues in IT301 and IT302. They provided a fun and vibrant environment to work.

Thanks to people at International Society of University of Manchester, they provided for me a fun environment to run the karate class. This gave me a great opportunity to meet people from different cultures. I feel blessed for sharing my karate experiences with them.

I would like to say thanks to my parents Nader and Nahid, and my lovely sisters Sepideh and Farinaz for their support and love. Without them I would not be what I am now.

Finally, I would like to say thanks, to my beloved husband, Mohsen, for his continues support, encouragement and patients without him this PhD can not be finished. To him I dedicated this thesis.

Dedicated to my husband

Chapter 1

Introduction

Vocal cords are the source of pressure variation for voiced speech (vowels) when the Bernouli effect [VdBZDJ57] causes them to periodically interrupt air-flow from the lungs. This pressure variation originates in the glottis, excites the vocal tract and propagates as voiced sound. Unvoiced speech (consonants) does not require vocal cord activity, though combinations of voiced and unvoiced speech occur frequently. Abnormalities in vocal cord tissue, for example due to inflammation, cause symptoms such as short term aperiodicity, breathy or hoarse voice and an inability to speak loudly.

The resulting loss of voice quality can be assessed objectively. Subjective or perceptual assessments of voice quality are commonly made according to a well-known standard referred to as 'GRBAS' [Hir81]. The GRBAS approach is widely used by Speech and Language Therapists (SLTs) in European hospitals and clinics. It was originally recommended by the Japanese Society of Logopeadics and Phonetics and a European Research Group [Hir81]. GRBAS is an acronym for five dimensions of voice quality referred to as 'Grade','Roughness', 'Breathiness', 'Asthenia' and 'Strain'. SLTs score their patients by giving a numerical value, i.e. an integer in the range 0 to 3, to each of these five dimensions of GRBAS. The approach is widely recognized and understood, but it is subjective, time-consuming and relies on highly trained human experts. Training SLTs in the GRBAS approach is expensive and time-consuming.

There are many algorithms in the digital signal processing (DSP) research literature for objectively measuring characteristics of voice from acoustical recordings. Such algorithms can distinguish voiced from unvoiced speech, measure the fundamental frequency of voiced speech, detect frequency or amplitude variation (jitter and shimmer), characterise the frequency spectrum in various ways and measure the extent to which voiced sound is affected by aperiodicity due to turbulent air-flow. Some of these algorithms are accessible as commercial or non-commercial software packages. Well known commercial packages are the Multi-Dimensional-Voice-Program (MDVP) [Kay96] and the Analysis of Dysphonia in Speech and Voice (ADSV) package [AR06]. A widely used non-commercial package is called 'Praat' (the Dutch word for 'talk') [Pa07]. The voice features measured objectively by these packages may be used to detect voice abnormality and may represent the characteristics of voice that SLTs listen for as indications of abnormality. However these DSP measurements are not GRBAS components and will not be familiar to non-DSP specialists who have wide experience with the GRBAS approach.

It is necessary to find a convergence between DSP based analysis and the perceptual evaluation of voice according to GRBAS. The aim will be design computerised systems able to make GRBAS voice quality assessments automatically [SLORGL$^+$08].

This thesis investigates the possibility of performing objective voice quality assessment conforming to the GRBAS scale. It considers how to make measurements, which voice features to measure and how to select features that produce the best possible predictors of GRBAS dimensions. A computerised objective version of GRBAS scoring could become a future standard for clinical use and research. There is currently no consensus for objective voice quality assessment according to GRBAS though there are many objective voice assessment schemes which have not been widely taken up for routine assessment. It is likely that the non-conformance of objective voice assessment systems to GRBAS is the main reason why they have not been widely adopted.

Figure 1.1 illustrates methodology of the project. A recorded voice signal will be fed into a digital system consisting of digital signal processing and mapping techniques based on machine learning. For each recorded voice sample, measurements of n voice features are made. In this work, 20 different features will be measured. Ten features will be measured using the commercial software package 'ADSV' and a different ten features will be measured by DSP algorithms developed specifically for this project. The 20 measurements, or specially selected subsets of them, will be supplied as input parameters, or 'features', to the mapping techniques.

To undertake this project, it was necessary to have access to a database of suitable voice recordings and to obtain GRBAS scores for each recording. Fortunately such a database, with scoring by five expert GRBAS scorers, was made available by MRI hospital. There are four main challenges. Firstly we must derive 'reference' GRBAS scores for each recording from the available data, and these must be made as reliable

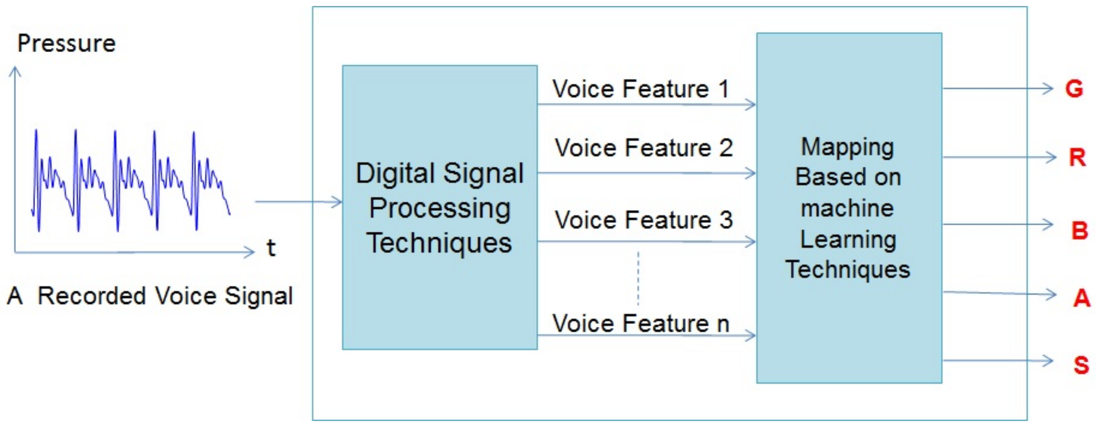


Figure 1.1: the goal of the project.

as possible. Secondly, we must discover which objective measurement features are associated with each of the GRBAS dimensions. Thirdly, we must find out which DSP techniques are appropriate for the feature measurement. Finally, we must decide which machine-learning algorithm can make the best prediction of GRBAS scores. A prototype objective system must be produced and evaluated by comparing its assessments against the reference GRBAS scores obtained from the trained SLTs.

1.1 Research Hypothesis

The research hypothesis is that computerised measurements of voice features processed using digital signal processing can be used, with machine learning, to produce GRBAS scores that are as useful and reliable as traditionally assessed subjective GRBAS scores

1.2 Aim and Objectives

The aim of this project is to design and evaluate methods for the objective measurement of voice quality conforming to the GRBAS standard with accuracy that matches that of trained SLTs.

The objectives are as follows:

1. To provide the means of obtaining GRBAS scores from trained SLT scorers for the database of voice recordings obtained at Manchester Royal Infirmary Hospital (MRI). This objective requires a scoring package (termed GPSP) which has features for archiving, checking and measuring the consistency of the scoring.

2. To devise a way of taking scoring consistency into account when producing the ‘reference’ GRBAS scores for training purposes.

3. To decide which voice features are most appropriate and how measurements of these features can be obtained reliably.

4. To determine how best to produce the necessary voice feature measurements using commercial and non-commercial DSP software and specially programmed DSP algorithms.

5. To verify the performance of the DSP algorithms being used by applying them to voice recordings with known characteristics and also by comparing the measurements they produce with similar measurements made by other software.

6. To discover a means of converting the DSP measurements to GRBAS scores using machine learning.

7. To investigate ‘feature selection’ techniques for finding the best subset of features for predicting the score for each GRBAS dimension.

8. To use the voice database, GRBAS scorings and statistical analysis to evaluate the results obtained by implementing a prototype objective GRBAS scoring system.

1.3 Research Contributions

In spite of some published research towards objective voice quality assessment according to the GRBAS scale, until now no definitive solution has been obtained. Some of the published approaches establish reasonable correlation between GRBAS components and voice feature measurements but do not progress to prototype objective systems that can do the GRBAS scoring automatically [BPG04]. Other approaches propose objective systems while not having the services of qualified GRBAS SLTs for producing the GRBAS scores needed for training their systems [VCOAAL$^+$13]. Published work does not address the issues of what is the appropriate number of features and what evaluation methods should be used [VCOAAL$^+$13]. The areas outlined below are the focus of this research contribution.

1. A major source of originality in this thesis is the use of the voice assessment database and the GRBAS scoring data established by MRI. The credit for this lies mainly with MRI.

2. The 'GPSP' software, and its graphical user interface, have novel features that, for example, allow scorer consistency and self-consistency to be assessed and taken into account when producing the reference GRBAS scores needed for training the prototype objective system. The intra-scorer consistency of each scorer which is assessed by requiring the scorer to repeat his/her assessment of a randomly selected subset of the subjects. The measurement of consistency and self-consistency using various forms of 'Kappa' is addressed, and a new form of the 'Fleiss Kappa', allowing it to be used for ordinal as well as categorical data, is proposed and investigated.

3. Although the DSP algorithms developed specifically for this project may not be fundamentally original, there is original insight in the decisions as to which methods to use for specific voice feature measurements, and the evaluation and comparisons with other published algorithms. In particular, the strong dependence of measurements harmonic-to-noise ratio on jitter and shimmer as obtained using the Praat software, and the fact that this dependence is greatly reduced by the software developed in this project, is an issue worth reporting.

4. The approach developed for objectively scoring Asthenia is original and has already been published. This is interesting because, although G, R, B and S properties have been well researched already, the assessment of Asthenia has been less extensively researched in the current literature.

5. The 'feature selection' methodology which finds the most appropriate feature subset for predicting each of the GRBAS components is original in its detail.

6. The idea of converting DSP measurements to GRBAS scores using K-Nearest Neighbor Regression (KNNR) and Multiple Linear Regression (MLR) is original.

1.4 Publications

1. Farideh Jalalinajafabadi, Chaitaniya Gadepalli, Mohsen Ghasempour, Frances Ascott, Jarrod Homer, Mikel Lujan, Barry Cheetham. 'Objective Assessment of

Asthenia using Energy and Low-to-High Spectral Ratio' 12 International Conference on Signal Processing and Multimedia Application, Alsace, France

2. Farideh Jalalinajafabadi, Chaitaniya Gadepalli, Mohsen Ghasempour, Mikel Lujan, Barry Cheetham and Jarrod Homer 'Computerised Objective Measurement of Strain in Voiced Speech' 37 Annual International Conference of the IEEE engineering in Medicine and biology society, Milan, Italy

3. Farideh Jalalinajafabadi 'Computerised Assessement of Overall Degree of Voice Abnormality' WomENcourage 2015, Uppsala, Sweden

4. Farideh. Jalalinajafabadi, Chaitanya.Gadepalli, Frances. Ascott, Jarrod. Homer, Mikel. Lujn, Barry. Cheetham, 'Objective voice quality assessment using digital signal processing and machine learning' UK Speech Conference, Edinburgh, United Kingdom, 2014

5. F.Jalalinajafabadi 'Automatic Analysis of GRBAS Scoring' in ACM womENcrourage Europe, Manchester, United Kingdom, 2014

6. F.Jalalinajafabadi, C.Gadepalli, F.Ascott, JJ.Homer, M.Lujn, B.Cheetham,'Perceptual Evaluation of voice Quality and its Correlation with Acoustic Measurement' in UKSim-AMSS 7th European Modeling Symposium on Mathematical modeling and Computer Simulation, Manchester, United Kingdom, 2013

7. C.Gadepalli, F.Jalalinajafabadi, F.Ascott, B.Cheetham, JJ.Homer 'Assessment of voice: subjective and objective measures' British Academic Conference in Otorhinolaryngology (BACO) 2015, Liverpool, Uk

8. C.Gadepalli, F. Jalalinajafabadi, F.Ascott, B.Cheetham, JJ.Homer 'Does subjective reflux reporting influence voice?' British Academic Conference in Otorhinolaryngology (BACO) 2015, Liverpool, Uk

9. C.Gadepalli, F.Jalalinajafabadi, F.Ascott, JJ.Homer, B.Cheetham 'Does VHI-10 correspond to GRBAS and CSID' BVA London,United Kingdom, 2014

10. C.Gadepalli, F.Jalalinajafabadi, F.Ascott, B.Cheetham, JJ.Homer 'Inter & intra rater consistency in GRBAS scoring' In Cutting Edge Larryngology, London, United kingdom, 2013.

1.5 Thesis Structure

There are seven chapters including this introductory chapter.

Chapter 2 explains the physiology of voice production and the approaches that may be used for voice quality assessment. It gives a description of perceptual subjective and computerised objective analysis. It also discusses the most important and recent techniques that have been published in the research literature for these two methods.

Chapter 3 describes our methodology in voice data collection and the application of GRBAS scoring by five SLTs to this database. Four different statistical methods, i.e. Pearson Correlation, Cohen's Kappa, Weighted Kappa and Fleiss' kappa are investigated for analysing the inter-scorer consistency and intra-consistency of the GRBAS scoring. A new approach for obtaining reliable 'reference' GRBAS scores is discussed in this chapter.

Chapter 4 discusses the DSP algorithms and commercial DSP tools that may be used for measuring the required acoustic features of speech. It surveys and examines the features that may be usefully measured and investigates the DSP mechanisms that are or may be employed to detect and quantify these features. Apart from ADSV, many published or commercialised techniques are applicable only to 'sustained' vowels, but it is explained how voiced speech may be extracted, for analysis, from 'connected' speech using the voiced/unvoiced decision provided by the 'thesis' software. The performances of these DSP algorithms are evaluated with reference to the MDVP and Praat software packages.

Chapter 5 explains the concept of machine learning and two approaches that may be applied to achieve objective GRBAS scoring. These approaches are Multiple Linear Regression (MLR) and K-Nearest-Neighbour-Regression (KNNR) . The objective scoring of each GRBAS component is based on measurements of twenty chosen acoustic features as identified in Chapter 4. Chapter 5 explores different methods for dimensionality reduction which may be expected to improve the performance of the objective scoring. Two different 'feature selection' method are focused on. These are 'filter methods' and 'wrapper methods'. These methods aim to identify the best feature subset for each GRBAS component. The performance of the objective scoring is evaluated in terms of Normalised Root Mean Square Error (NRMSE) and Pearson correlation.

Chapter 6 evaluates the results obtained for the objective scoring of each GRBAS component. A prototype objective scoring system is evaluated against the 'reference' GRBAS scores. Results are presented showing the performance of prediction models ‘with feature selection’ and ‘without feature selection’. All twenty features are first used for each GRBAS component and then the best subset of features obtained using feature selection is used.

Chapter 7 gives conclusions and suggestions for further work on this research topic.

Chapter 2

Background and Related Work

2.1 Introduction

This chapter gives background details about speech, the human speech production mechanism and related work in voice quality analysis. The causes and common manifestations of voice impairment are discussed. The chapter then surveys how voice problems can be investigated by the perceptual evaluation of voice features performed by the person himself and/or by a clinician. Two standardized ways of performing and recording the results of clinical evaluations are reviewed; these are the GRBAS approach as widely used in Europe and the CAPE-V approach developed in the USA. Self-assessment using a standardised questionnaire known as VHI-10 (Voice Handicap Index) is discussed as a useful adjunct to clinical assessment. Then the chapter discusses the objective analysis of speech and the possibility of augmenting or replacing subjective clinical analysis by objective computerised methods. Much has been done already in this field, but up to now the results of objective analysis have not been presented in a form that is recognizable to clinicians. The aim of producing assessments conforming to the GRBAS scale is now discussed in more detail. The literature on this topic is surveyed. Digital signal processing techniques that are available for measuring certain parameters of speech are introduced and analysed. The need for further refinement to published and commercialised algorithms is discussed, and the ideal of analysing connected speech as well as sustained vowels is explained. Since the objective assessment techniques to be developed will be based on a number of machine learning techniques, some background on these techniques will be presented.

2.2 Speech and human speech production

Speech is sound which is a variation in air pressure. It is conveyed as a wave which travels through air at about 34320 cm/s. Human speech production requires a flow of air forced out from the lungs by the breathing mechanism. Without any subsequent modulation, this air-flow would produce only sound due to turbulence which would sound like a 'random signal' without any information. Random signals are created by waterfalls, the sea, badly tuned AM radios, cars traveling at speed and many other every-day effects, The sound produced is often described loosely as 'random noise' and assumed to be 'white' which means evenly spread over a wide spectrum of frequencies. In speech, the turbulence can be modified in volume and frequency spectrum by creating a constriction at the back of the throat Arabic /h/, at the centre or front of the mouth (/sh/ and /s/), at the teeth and lips /f/, explosively at the lips /p/ and elsewhere. Such sounds are termed 'unvoiced' and created consonants which tend to be transient but actually carry most of the information within speech.

Vowel sounds require an extra mechanism which modulates the air-flow from the lungs in a different way. This mechanism is performed by the 'vocal folds' which are also called 'vocal cords' and reside in the larynx sometimes known as 'Adams apple'. The vocal folds are highly elastic muscular tissue which in normal people can close completely (or almost completely) to momentarily interrupt the air supply from the lungs. This closure builds up pressure behind the vocal cords that eventually forces them open. The pressure then reduces again and the vocal cords can close once more. This mechanism creates a pressure variation which is close to being periodic. Most energy is created when the vocal cords 'snap' closed, and this happens between about 80 and 160 times per second in adult talkers. The almost periodic pressure variation created by the closing vocal cords determines the pitch of the voice in speaking. The frequency of the vocal cord variation is termed the 'fundamental frequency' F_0 of voiced speech. Children have much higher values of F_0 than adults, and singers use the same mechanism as talkers but over a wider F_0 frequency range.

The sound produced by the vibrating vocal cords is often modeled as a periodic series of impulses doubly integrated (approximately) by a low-pass 'glottal filter' whose gain response resembles that of a digital filter with two poles. The spectrum of a periodic series of impulses, expressed as a Fourier series, has a fundamental sinusoid at the fundamental frequency of the vibrations, plus a series of harmonics at twice, three times, four times the fundamental frequency and so on extending to infinity in theory. In practice, the harmonics become very small beyond about 4 kHz, but there may be as many as 50 harmonics that affect the sound significantly. Without the glottal filter, the amplitudes of the harmonics would in principle be all equal. The glottal filter imposes a loss which increases with frequency at about 12 dB per octave above about 100 Hz. This is the signal that would be detected at the site of the vocal cords by an 'electro-glottogram'(EEG) [CHMA86, Nat14]. It does not sound like a vowel. However it serves as an excitation signal to the vocal tract which comprises the mouth and nasal cavities.

The excitation signal is now modulated spectrally and in amplitude by the resonances of the vocal tract and the effects of the nasal cavities and the teeth and lips. The true vowel sound is thus produced and is affected by the shape of the mouth as controlled by the jaw, tongue, teeth, lips and velum (which connects or disconnects the nasal passage). Figure 2.1 shows a DSP model of the process described above, often referred to as the 'source-filter' model of speech production. In this model, the vocal tract is represented by a time varying digital filter whose frequency response determines the phonetic content of the sound. For voiced sounds, H(z) is often a tenth order all-pole filter whose poles produce the required vocal tract resonances. For unvoiced sounds, H(z) also controls the amplitude of the output including the rapid transitions that occur with consonants in speech. The 'gain' constants A_v and A_{uv} control the overall amplitudes of the voiced and unvoiced components (respectively) of the speech. A more detailed description of the vocal folds, glottal area and vocal tract is given in the next section.

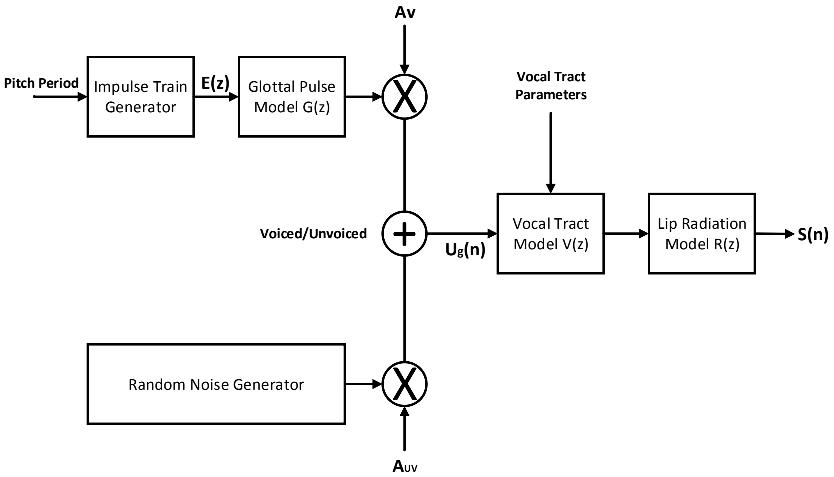


Figure 2.1: Speech Production Model

2.2.1 Vocal folds

The vocal folds are an essential part of the human anatomy for voiced speech production. They are located in the larynx and also serve to protect the air-ways from choking on material in the throat. During respiration, the vocal folds are abducted and they allow airflow from the lungs to move freely in and out of the body. Forming constrictions in the airway is a critical part of phonation and, in fact, the vocal cords can intrude dynamically on the air-stream due to the Bernouli Effect [VdBZDJ57]. This effect occurs when passing a gas or fluid with constant flow through a tube when a section of the tube is constricted. At the point of constriction the flow will speed up and there will be a drop in pressure against the walls of the constricted part of the tube. Where the constriction is caused by the vocal cords, their mechanical properties, the length, the thickness, the tension and the vibrating mass of the vocal folds determine how they react to the fall in pressure. The properties of the vocal cords are controlled by the larynx which contains cartilages and muscles whose properties may be varied by the human talker.

The vocal cords are composed of layers of soft tissue where each layer has different properties. Each layer has a degree of elasticity and is capable of some independent movement. The air-stream is interrupted by the vocal folds, when they are adducted during phonation. At this point, sub-glottic pressure begins to build up below the vocal cords. The pressure eventually forces the soft tissue to separate and the air-stream is then allowed to flow through the vocal cords again. According to the Bernouli Effect, when the air-stream through the vocal folds accelerates, a drop in pressure occurs which causes the vocal cords to come back together. Sub-glottic pressure then builds up again and the process continues. This process of vocal fold motion creates the air-pressure compressions and rarefactions from which all the vowels can be generated.

2.2.1.1 Vocal fold vibration measurement

Vocal folds vibrate at a frequency which becomes the fundamental (pitch) frequency of voiced sound. The length, tension and mass of the vocal cords, and the sub-glottal pressure created are four factors which determine the frequency and nature of the vibration. Irregular cycle-to-cycle variation in the vibration can be a natural characteristic of some people's voices, though it may also be caused by different kinds of voice disorder. Observing the vocal fold vibration can be very useful for recognizing voice disorders. Invasive methods involve the insertion of tools like catheters and balloons into the body for the treatment and diagnosis while non-invasive methods involve imaging by ultrasound and nuclear tracer imaging. Measuring the vocal fold vibration directly must use a non-invasive method because the larynx is clearly not easily accessible during the phonation. Researchers have produced several non-invasive methods for directly monitoring vocal folds vibration. Electro-glottograph (EGG) and acoustic measurement are two non-invasive methods that do not require surgery or internal examination. They can obtain measurements of vocal fold vibration in real time while a person is speaking.

An EGG monitors the variations of electrical impedance across the larynx. It measures variations in electrical conductance by applying a small potential difference between two electrodes placed on the throat and measuring the variation of current that occurs. When the vocal folds are closed, the impedance decreases and more current flows than does when they are open. However it is not always desirable or convenient to connect electrodes to a patient. The EGG device is not very comfortable and is often not very reliable. It requires training for an investigator to be able place the electrodes correctly, and the change of impedance may be reduced and difficult to detect in patients with excessive neck fat. Figure 2.2 shows an EGG device applied to a patient's larynx. EGG waveforms were obtained as part of the procedure used by Gallepalli [C.G13a] to establish the data-base used in this thesis. This is not intended to be part of the standard procedure to be adopted for examining patients in future. It was done as a research tool for verifying the results obtained from the DSP analysis of purely acoustic signals. Figure 2.3 shown the EGG waveform obtained from a patient for the short 45.35 ms segment of the sustained vowel /a/ shown in Figure 2.4.

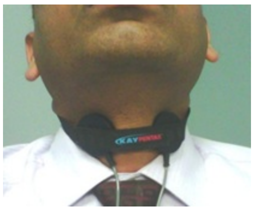


Figure 2.2: EGG device applied to a patient's larynx

The sampling frequency is 44100 Hz. It may be noted that the impedance reduces relatively slowly during the vocal cord opening phase of each pitch cycle, and then increases sharply as the vocal cords 'snap' together to initiate the closed phase. This waveform was obtained using an EGG with the commercial DSP tools supplied by Kay-Pentax [Kay96]. These tools apply unspecified processing to obtain ‘clean’ waveforms and it can be seen that this processing introduces some delay.

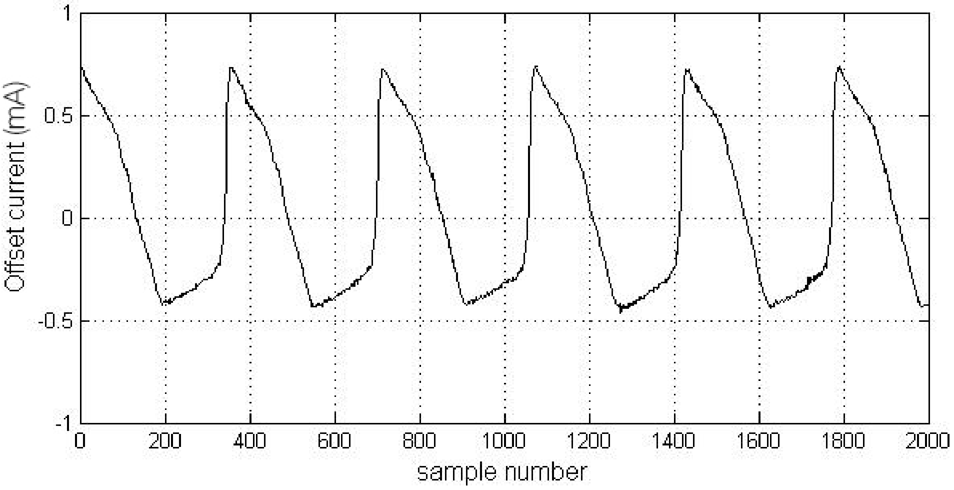


Figure 2.3: EGG waveform produced by short 45.35 ms segment of the sustained vowel /a/. The sampling frequency is 44100 Hz

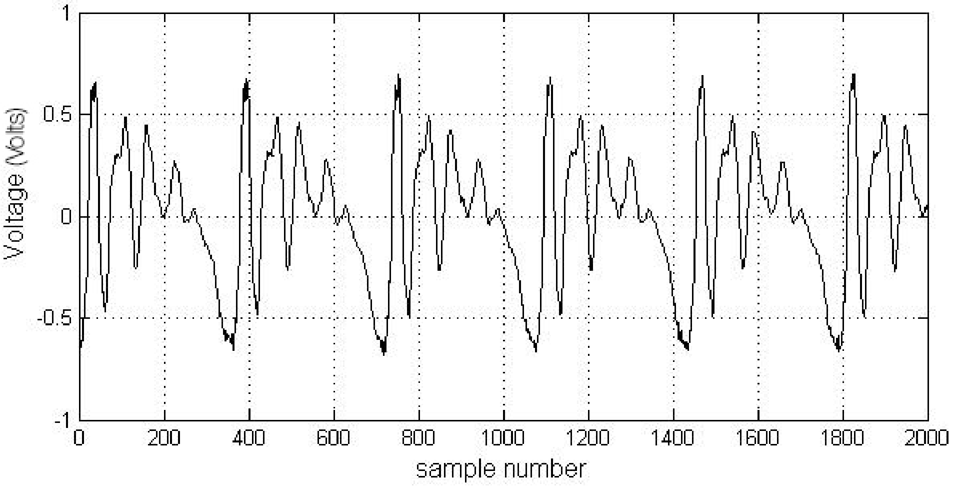


Figure 2.4: Speech waveform produced by short 45.35 ms segment of the sustained vowel /a/. The sampling frequency is 44100 Hz

2.2.2 Vocal tract

The vocal tract provides the airway used in the production of speech. It includes the throat, mouth, palate, tongue, teeth and lips. The nasal passage provides a coupled airway which contributes marginally to the production of many sounds and is used exclusively or largely for some voiced ‘nasal’ phonemes such as /m/ and /n/. For voiced speech, the vocal tract is responsible for changing the spectral balance of the glottal source signal and modulating its amplitude, for example by closing the airway using the lips or tongue. Formants are peaks in the spectral envelope of voiced speech. They are superimposed on the flat spectrum of the glottal excitation signal by resonances of the vocal tract. The frequencies and Q-factors of these resonances change according to the shape of the vocal tract during speaking. Talkers can make a wide variety of voiced sounds by changing the vocal tract shape. The resonances (and anti-resonances) also affect unvoiced speech and the vocal tract is responsible for the constrictions that produce the turbulence responsible for the unvoiced sound. For example the position of the constriction created by the tongue and the roof of the mouth, and the resonance created by the vocal tract shape thus formed are responsible for the consonants /sh/ and /s/ and the difference between them. On average, the total length of the vocal tract from the larynx to the lips/nostrils is about 17-18cm (in men) [Fit97]. If the tube were open and straight, its length and the speed of sound c = 34320 cm/s would determine the frequency of the lowest resonance at around $c/(2\times 18) = 953$ rad/s or 152 Hz because the maximum wavelength of a pitch-cycle would be about $2\times 18 = 36\ cm$. A closed tube can resonate at the same frequencies as an open tube of twice its length. The more complex shape of the vocal tract can create a range of formants which can be varied by moving tongue and lips, and even stretching the length of the vocal tract, and this variation make different speech sounds. There are normally about three to five observable formants in voiced sounds depending on the phoneme sound and the speaker characteristics.

2.2.3 Speech phonemes

A phoneme is the smallest acoustic element of a spoken word that can change its meaning; for example /c/ , /b/ and /e/ which can change ‘cat’ to ‘bat’ or ‘bet’. Phonemes can be classified in many ways, but is it useful to have three categories

1. Voiced phonemes which are approximately periodic over short time periods and can have a fundamental frequency in the range 80 to 600 Hz. Fundamental frequencies at the extremes of this range are rare but possible. During normal speech, the fundamental frequency depends on different factors such the sex of the speaker, age, intonation and emotional context. Voiced phonemes are

produced from quasi-periodic pulses of breath which excite the vocal and they include phonemes which are labeled /U/,/d/,/w/,/i/ and /e/.

2. Unvoiced sounds or fricatives are produced by forming a constriction in the vocal tract. The air-flow forces breathe through the constriction at high velocity to generate turbulence which produces ‘noise-like’ sound that excites the vocal tract. These sounds include phonemes which are labeled [s], [z], [ʃ].

3. Plosive phonemes are voiced or unvoiced sounds produced during the explosive release of air-pressure following a complete closure formed by the tongue or lips. Examples are labeled as t, k, and p (voiceless) and d, g, and b (voiced).

2.3 Voice disorder

Deciding what characterises an abnormal voice is difficult and beyond the aims of this thesis. A normal person’s voice is individual and often immediately recognisable, sometimes with characteristic breathiness, roughness, frequency and amplitude variations and other features that are also of interest in studying pathological voices. It must be assumed that decisions about which voices are disordered have already been taken by clinicians and this thesis is concerned only with investigating and quantifying the effects of the disorder, assuming it exists. The voice of an individual will change from day to day depending on many factors such how it has been used recently, fatigue, illness and the person’s emotional state. The voice may change from morning to night without obvious reason. It is generally a severe change beyond these normal variations that triggers an investigation.

Two temporary causes of voice disorder are vocal cord misuse and inflammation associated with common colds and referred to as laryngitis. More serious and longer term causes include vocal fold paralysis, polyps, nodules and cancer. Signs of voice disorder are usually observed and first mentioned by the patients [CL06]. These signs include pain in the throat or larynx due to inflammation, abnormal pitch, breathiness, uncontrolled variation in amplitude and fundamental frequency, insufficient loudness and other unwanted changes in the quality of the sound produced by the larynx. Voice problems can have a negative effect on the quality of life of those who suffer from them [MY01, Yiu02, RMT$^+$04]. Careers as well as day to day activities can be adversely affected. Our experience of working with Dr. Gadepalli at MRI Hospital has confirmed this beyond doubt. Hoarseness is a commonly used term for people who have voice disorder. If somebody has a hoarse voice the sound is breathy, irregular, constantly changing, sometimes absent altogether, and often softer in volume and/or lower in pitch. Hoarseness is often a symptom of problems in the vocal folds of the larynx, though a degree of roughness (like 'gravel'), which may be described as hoarseness, is a natural and cherished feature of some voices.

Laryngitis, often associated with the 'common cold' and flu, causes hoarseness, roughness, and pain making it difficult or impossible to speak. A person 'loses his/her voice'. The nature of the problem seems to change from moment to moment. This is because of a constantly changing accumulation of mucous around the vocal folds. This mucous is a colourless viscous fluid secreted by certain cells in the larynx. It is oily and sticky and serves to lubricate the vocal folds when they are working normally. It is produced in excess when the vocal cords are inflamed, and similar mucous produced by the lungs may be deposited on the vocal cords when the chest is inflamed during a cold or flu. The consistency of the mucus on the vocal cords may also change, making it less effective as a lubricant.

The inflammation and excessive build up of mucous make the vocal folds heavier and less elastic. They are no longer able to 'snap' closed quickly with the energy needed to produce a vigorous periodic excitation with strong harmonic content. They may fail to close completely either consistently or sporadically. The vibration may reduce in frequency resulting in a deeper voice, and the lack of harmonic content may produce dull or muffled sounding vowels. The mucous no longer forms an even film over both of the vocal cords as required to cause them to come together evenly and securely. Consequently, air escapes through partially closed vocal cords in a random and unpredictable manner and even 'bubbles through the mucus'. The result is aperiodic, breathy, uneven and unpredictable voiced sound associated with hoarseness.

The absence of mucus can also be a problem since it helps the vocal cords to close evenly and completely. A dry throat can produce aperiodic, breathy speech which can often be alleviated by a drink of water.

Vocal cord misuse occurs when a child in the playground or an adult at a football match shouts too loudly and for a long time. Football managers frequently bear the consequences of such misuse. A similar problem occurs with amateur choral singers who sing too loud and enthusiastically and lack the skill of professional singers to protect their voices. During shouting or loud singing the vocal folds come together with much greater force than is required for normal speaking. This irritates the vocal folds and interferes with the normal function of the lubricating mucous. The vocal cords can become dry and inflamed, and further speaking can become very painful and sometimes impossible for a while.

Fortunately, the difficulties mentioned above are usually temporary, though there is evidence that frequent and prolonged misuse of the voice by some professionals, such as teachers [C.G13b], can have long term and even permanent effects. There is also evidence that changes in voice characteristics may be the first sign of a serious voice disorder[CL06, SGG00].

More serious voice disorders with longer term causes include vocal fold paralysis, polyps, nodules and cancer. These can be very serious.

2.4 Voice quality analysis

The term ‘voice quality’ often refers to the perceptual characteristics of a voice as heard by another human being. This may not be an ideal notion since humans may not be as sensitive to certain abnormalities as an objective analysis. It is argued by Jody et al. [KG03] that although jitter and shimmer analysis are ‘the cornerstone of acoustic voice measurement’, they are not very noticeable to humans. The use of perceptually-biased parameters, such as mel-scale Mel Frequency Cepstral Coefficients (MFCCs) as in reference [MBE10], must be questioned since they exploit human perception to reduce the information content of the signals they represent. What a human being will not notice is not recorded. However, it is possible that the missing information is, after all, significant. Maybe not, but it is best not to take the risk.

This thesis is primarily interested in analysing phenomena arising from the phonatory action of the laryngeal system [LKB00]. The perceived grade or quality of hoarse pathological voice is often described in terms of roughness, breathiness, weakness and the apparent strain involved in producing the sound. These properties are the basis of the ‘GRBAS’ method of perceptual grading which is so well known in Europe that it is desirable to base any computerised objective method on GRBAS. There are some different methods used for voice quality assessment, both subjective and objective, and these will be surveyed later. The extent to which computerised objective methods can reproduce perceptual grading will be investigated, but it must be remembered that a computer method can notice aspects of speech that a person may miss. It is also probably true that visual cues may be helpful to the clinician and not available to the computer program. For example, visual cues may be especially helpful in assessing the strain needed to produce a sound. The next section is concerned with the GRBAS assessment method [Hir81].

2.4.1 Voice Handicap Index

Voice Handicap Index (VHI) was introduced by Jacobson et al. [JJG$^+$97]. It is a self-assessment of voice quality by participants who attend voice clinics. They fill out 30 items in a questionnaire and describe their voice and its effect in their life. The questionnaire covers different aspect of voice disorder such as functional, physical and emotional aspects. VHI-10 requires little time to administer and is easy to score. The VHI with 30 items is evaluated with the following range of responses: 'never', 'almost never', 'sometimes', 'almost always', and 'always'. In analysing the questionnaire, each response was scored from 0 to 4. The, total score could range from 0 to 120 points. Figure 2.5 is an VHI-10 form.

2.4.2 GRBAS

GRBAS is a scheme for voice quality assessment based on a multidimensional analysis of perceived voice qualities. It evolved from the work of several researchers and was popularised after being described by Hirano in 1981[Hir81]. GRBAS stands for the five assessments required, which are Grade, Roughness, Breathiness, Asthenia, and Strain. The GRBAS scale is considered to be the major and most reliable tool [WCD$^+$04] in perceptual speech quality evaluation. Physicians and 'Speech and Language therapists' use it routinely to assess patients and monitor their progress before and after therapy. The five 'components' or attributes of GRBAS are the descriptors of perceived voice quality. For each attribute a four-point scale is used to specify the severity of any perceived abnormality: '0' indicates none, '1' signifies slight abnormality, '2' signifies moderate abnormality and '3' signifies severe abnormality. The scale may be considered ordinal, with magnitudes ordered in terms of increasing severity.

Grade (G), represents the overall degree of hoarseness or voice abnormality.

Roughness (R) quantifies the degree to which the listener detects the effect of irregular fluctuations in pitch-frequency and amplitude either cycle to cycle or in the short term energy of the vocal tract excitation [Hir81]. Roughness is also affected by perceived randomness or 'noisiness' of the spectrum [HMWM66]. Any perception of roughness might take into account the possibility of severe irregularity due to vocal fry [Hir81, HMWM66] and double excitation (diplophonia) [Sch95].

Breathiness (B) arises from non-periodic sound generated by a turbulent flow of

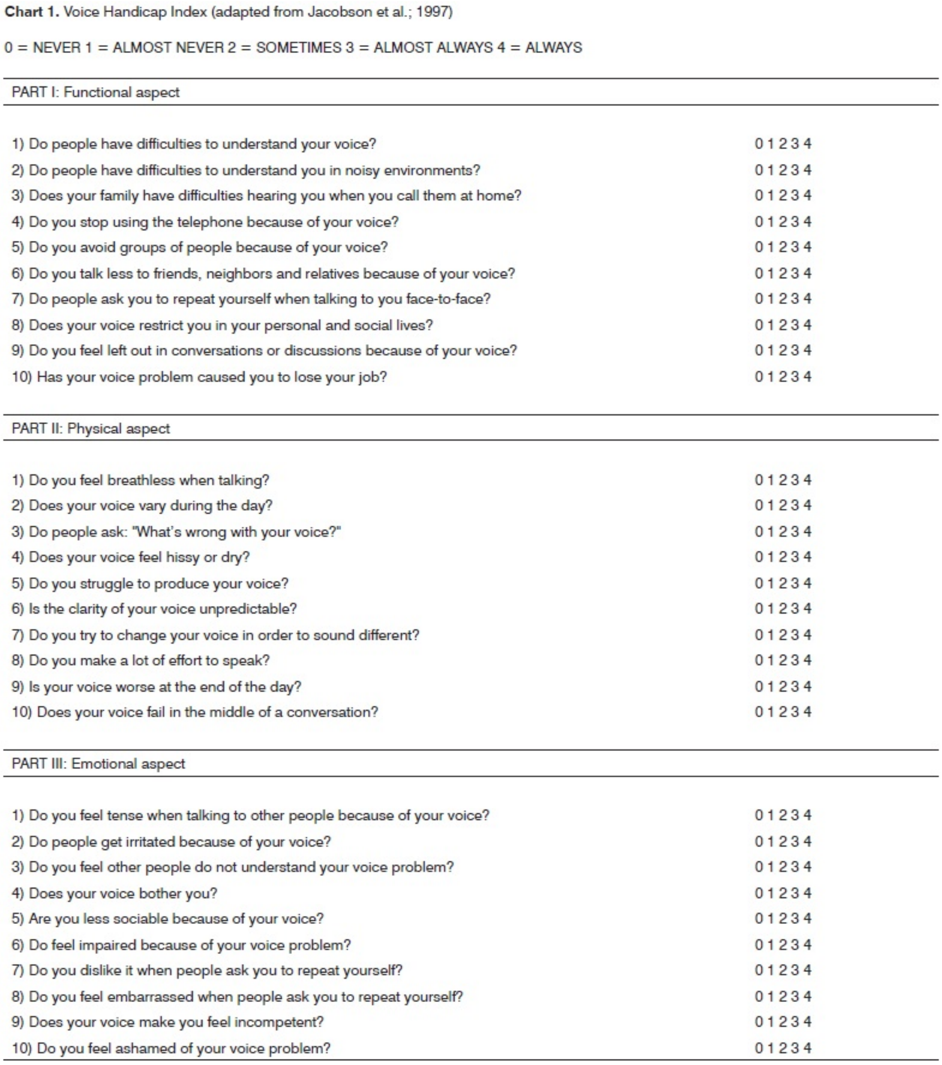


Chart 1. Voice Handicap Index (adapted from Jacobson et al.; 1997)

0 = NEVER 1 = ALMOST NEVER 2 = SOMETIMES 3 = ALMOST ALWAYS 4 = ALWAYS

PART I: Functional aspect	
1) Do people have difficulties to understand your voice?	0 1 2 3 4
2) Do people have difficulties to understand you in noisy environments?	0 1 2 3 4
3) Does your family have difficulties hearing you when you call them at home?	0 1 2 3 4
4) Do you stop using the telephone because of your voice?	0 1 2 3 4
5) Do you avoid groups of people because of your voice?	0 1 2 3 4
6) Do you talk less to friends, neighbors and relatives because of your voice?	0 1 2 3 4
7) Do people ask you to repeat yourself when talking to you face-to-face?	0 1 2 3 4
8) Does your voice restrict you in your personal and social lives?	0 1 2 3 4
9) Do you feel left out in conversations or discussions because of your voice?	0 1 2 3 4
10) Has your voice problem caused you to lose your job?	0 1 2 3 4
PART II: Physical aspect	
1) Do you feel breathless when talking?	0 1 2 3 4
2) Does your voice vary during the day?	0 1 2 3 4
3) Do people ask: "What's wrong with your voice?"	0 1 2 3 4
4) Does your voice feel hissy or dry?	0 1 2 3 4
5) Do you struggle to produce your voice?	0 1 2 3 4
6) Is the clarity of your voice unpredictable?	0 1 2 3 4
7) Do you try to change your voice in order to sound different?	0 1 2 3 4
8) Do you make a lot of effort to speak?	0 1 2 3 4
9) Is your voice worse at the end of the day?	0 1 2 3 4
10) Does your voice fail in the middle of a conversation?	0 1 2 3 4
PART III: Emotional aspect	
1) Do you feel tense when talking to other people because of your voice?	0 1 2 3 4
2) Do people get irritated because of your voice?	0 1 2 3 4
3) Do you feel other people do not understand your voice problem?	0 1 2 3 4
4) Does your voice bother you?	0 1 2 3 4
5) Are you less sociable because of your voice?	0 1 2 3 4
6) Do feel impaired because of your voice problem?	0 1 2 3 4
7) Do you dislike it when people ask you to repeat yourself?	0 1 2 3 4
8) Do you feel embarrassed when people ask you to repeat yourself?	0 1 2 3 4
9) Does your voice make you feel incompetent?	0 1 2 3 4
10) Do you feel ashamed of your voice problem?	0 1 2 3 4

Figure 2.5: VHI-10 voice is a self-assessment of voice quality by participants who attend voice clinics. The VHI with 30 items is evaluated with the range of responses: ‘never’, ‘almost never’, ‘sometimes’, ‘almost always’, and ‘always’. In analysing the questionnaire, each response was scored from 0 to 4. The, total score could range from 0 to 120 points.

air which leaks through the glottis when it is supposed to be closed [Hir81]. The turbulence is created by the constriction of a partially closed glottis. Its energy will be correlated to the vocal cord activity; i.e. its energy will decrease as the glottis becomes fully open and increase again as the vocal cords try to close. At its source, the turbulence will be spectrally flat (white) but it will spectrally coloured by the vocal tract resonances and maneuvers (e.g. opening/closing at the lips) as it contributes to perceived speech. As the sound heard from normal breath or unvoiced speech is due to turbulent air-flow caused by some constriction in its passage, the sound created by imperfectly closing vocal cords will sound like breath or unvoiced speech. The perceived quality of breathy voice quality is related to the amount of air-flow. Breathy voice lacks clarity of tone and is reduced in loudness. Most voices have a degree of breathiness which contributes to their individuality and natural characteristics.

Asthenia (A) is weakness or lack of energy in the voice. The asthenic variety of hoarse voice is mostly characterised by weak intensity [Hir81]. It can be because of an impaired energy distribution in the glottal excitation with a spectral damping which is a sign of a lack of elasticity in the vocal cords. The higher harmonics in the perceived sound will then have a lack of brightness and richness.

Strain(S) is indicative of undue effort needed to produce voiced sound due to an inability to employ the normal functionality of vibrating vocal cords [Hir81]. There is often psychological stress involved in trying to overcome the disability and this is perceivable by the trained listener. The abnormally functioning vocal cords and the stress in trying to control them can produce sound with abnormally high fundamental frequency, with unnatural and constantly changing periodicity and roughness in the higher frequency range of the speech. Strain due to speaking with abnormality functioning vocal cords is perhaps the most subjective GRBAS measurement and the most variable effect. Strain is associated with increased and poorly regulated laryngeal muscle tension [Hir81, CL06]. When speech is being produced, there is the perception of an inability to control it as it fades in and out. Difficulty in initiating phonation and a struggle to maintain phonation takes place due to strain. Furthermore, constantly changing periodicity in the higher frequency harmonics is indicative of strain, giving the perception of noise or roughness in the higher frequency range of the speech.

2.4.3 Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)

The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) was developed as a clinical tool for perceptual assessment of voice by the American Speech-Language-Hearing Association's (ASHA) Division 3 [KGA$^+$09]. They developed standardised guidelines for the perceptual assessment of voice based on voice perception, psychometric scaling and data in psychoacoustics. CAPE-V is an initial product that, it is hoped by some, will soon be in widespread use by clinicians, SLTs and researchers. The CAPE-V standard measures important perceptual vocal attributes and is intended to be easily understood and used by many professionals. The properties it measures are overall severity, roughness, breathiness, strain, pitch, and loudness. It is therefore an alternative to GRBAS, but includes two additional measurements and does not reflect the perception of asthenia in the way GRBAS defines it.

Overall severity arises from integrated voice deviance. Roughness is described as perceived irregularity in the voicing source. Strain is described as perception of excessive vocal effort. Pitch is a perceptual correlate of pitch-frequency (fundamental frequency) which is scored by clinicians according to how they consider the fundamental frequency of the subject to deviate from what may be expected for the subject's age, relevant culture and gender.

Loudness is a perceptual correlate of sound intensity which is intended to reflect the deviation of the subject's voice from normal for a person of his/her age, referent culture and gender.

A graphical user interface (GUI) in Figure 2.6 supplied by Kay-Pentax [Kay96] represents CAPE-V attributes by a 100-millimeter line forming a visual analog scale (VAS).

Entering scores may be assisted by referring to general regions indicated below each scale. 'MI' refers to 'mildly deviant', 'MO' refers to 'moderately deviant' and 'SE' refers to 'severely deviant'. Clinicians tend to use these regions to indicate the severity of any degradation, rather than the numerical scale. Scorings are based on direct assessment, by a trained clinician, of the subject's performance with vocal maneuvers. Figure 2.7 is an illustration of CAPE-V form. CAPE-V assessments are sometimes augmented by self-assessments as provided by a completed Voice Handicap Index (VHI) questionnaire or a similar self-assessment [JJG$^+$97].

The CAPE-V assessment requires each attribute to be classed as either 'consistent' (C) or 'intermittent' (I). A 'consistent' classification means that the attribute was continuously observed throughout the assessment. An 'intermittent' classification means that the attribute occurred inconsistently within or across the assessment. For instance, a subject may consistently exhibit a breathy voice quality across all the assessment, which includes sustained vowels and speech. On other hand, the subject might exhibit consistent breathiness during vowel production, but intermittent breathiness during one or more connected speech phonation. The scorer would then classify the breathiness as intermittent. CAPE-V is a widely used technique in the USA while GRBAS is mostly used in Europe and other continents.

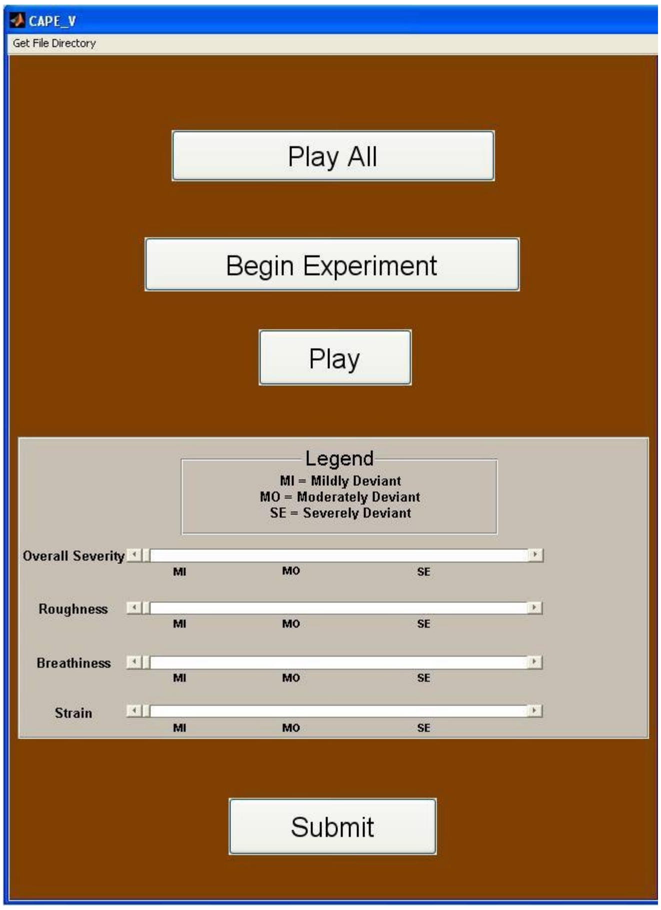
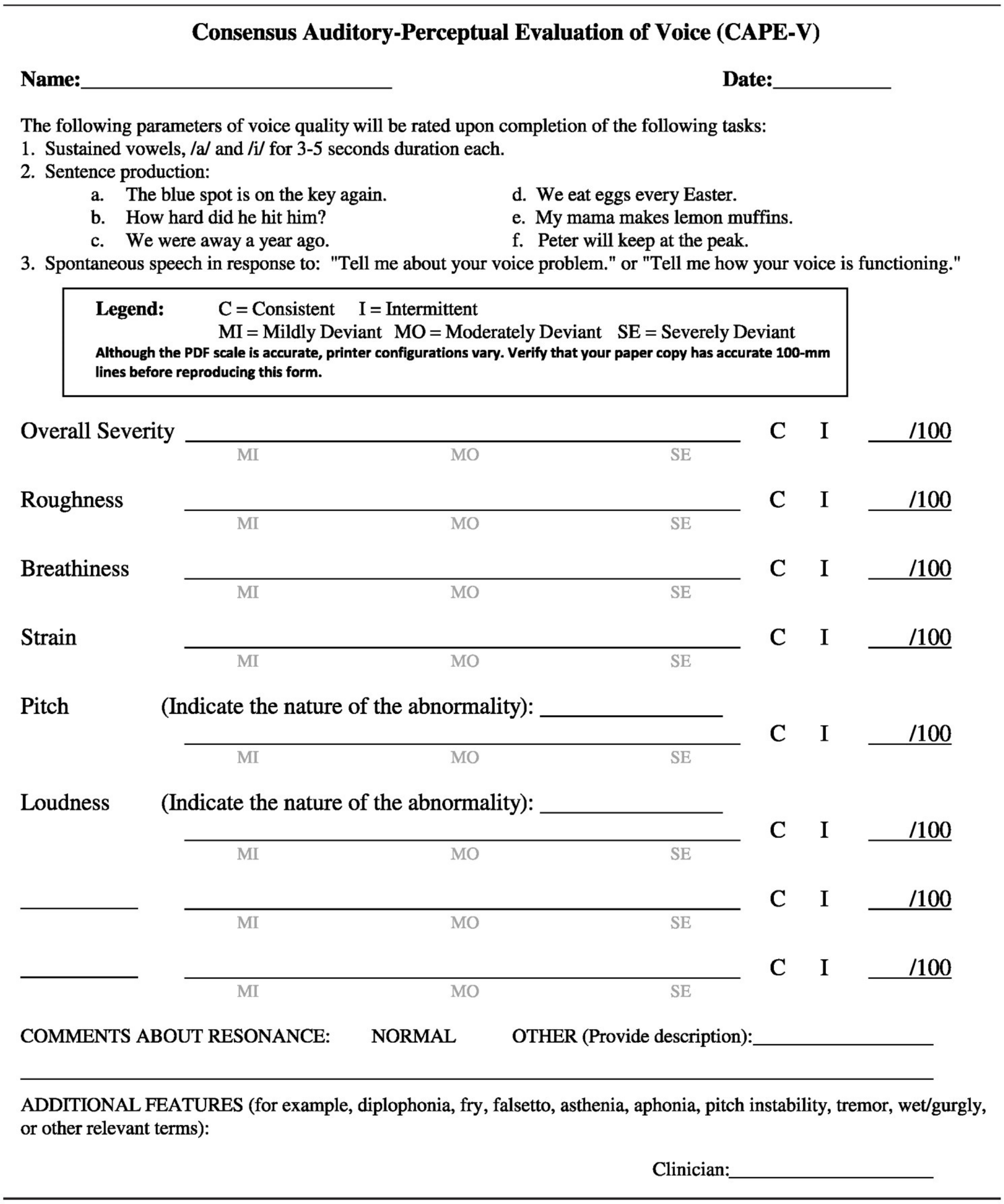


Figure 2.6: A graphical user interface (GUI) in supplied by Kay-Pentax [Kay96] represents CAPE-V attributes (Overall severity, Roughness, Breathiness, Strain) by a 100-millimeter line forming a visual analog scale (VAS).‘MI’ refers to ‘mildly deviant’, ‘MO’ refers to ‘moderately deviant’ and ‘SE’ refers to ‘severely deviant’.

2.5 Objective voice measurement

The purpose of objective measurement is to use computerised measurement techniques to analyse the quality of a person’s voice. Objective voice measurement can be performed on acoustic recordings of vocal maneuvers consisting, typically, of sustained vowels and passages of connected speech captured by a suitable microphone. Acoustic recording of connected speech are preferred for diagnostic purposes, although they can be more complex to analyse. The analysis of connected speech normally requires voiced/unvoiced decisions to identify voiced and unvoiced sections of the speech. Subsequent analysis is then applied, normally, to just the voiced sections.

Voiced speech produces the vowels in continuous speech, whereas unvoiced speech



Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)

Name:____________________________ **Date:**___________

The following parameters of voice quality will be rated upon completion of the following tasks:

1. Sustained vowels, /a/ and /i/ for 3-5 seconds duration each.
2. Sentence production:
   - a. The blue spot is on the key again. d. We eat eggs every Easter.
   - b. How hard did he hit him? e. My mama makes lemon muffins.
   - c. We were away a year ago. f. Peter will keep at the peak.
3. Spontaneous speech in response to: "Tell me about your voice problem." or "Tell me how your voice is functioning."

Legend: C = Consistent I = Intermittent
MI = Mildly Deviant MO = Moderately Deviant SE = Severely Deviant
Although the PDF scale is accurate, printer configurations vary. Verify that your paper copy has accurate 100-mm lines before reproducing this form.

Overall Severity ______________________________ C I ___/100
MI MO SE

Roughness ______________________________ C I ___/100
MI MO SE

Breathiness ______________________________ C I ___/100
MI MO SE

Strain ______________________________ C I ___/100
MI MO SE

Pitch (Indicate the nature of the abnormality): _______________
______________________________ C I ___/100
MI MO SE

Loudness (Indicate the nature of the abnormality): _______________
______________________________ C I ___/100
MI MO SE

__________ ______________________________ C I ___/100
MI MO SE

__________ ______________________________ C I ___/100
MI MO SE

COMMENTS ABOUT RESONANCE: NORMAL OTHER (Provide description):__________________
____

ADDITIONAL FEATURES (for example, diplophonia, fry, falsetto, asthenia, aphonia, pitch instability, tremor, wet/gurgly, or other relevant terms):

Clinician:____________________

Figure 2.7: This is the CAPE-V form. Clinicians measures overall severity, roughness, breathiness, strain, pitch, and loudness out of 100.'MI' refers to 'mildly deviant', 'MO' refers to 'moderately deviant' and 'SE' refers to 'severely deviant'.

produces the consonants. Voiced speech sections are identifiable as having quasi-periodic pressure waveforms whereas unvoiced sections do not have this quasi-periodicity. Quasi-periodic means that a fundamental cycle can be identified, within the speech waveform, which is repeated approximately for several consecutive cycles. Within these repeated cycles, there may also be non-periodic components such as additive noise. The approximate nature of the periodicity means that the fundamental frequency, the amplitude, and the characteristic shape of the voiced speech can change over time.

The fundamental frequency (referred to as F_0) determines the ‘pitch’ of the voice which is typically around 80 Hz to 180 Hz for male speech and around 165 Hz to 255 Hz for female speech. In normal voiced speech, there should be strong localised periodicity, relatively little uncontrolled amplitude or fundamental frequency variation from cycle to cycle, and there should also be a relatively little random ‘noise-like’ component due to turbulent air-flow. Abnormal voiced speech, produced by damaged vocal cords, is characterised by much higher uncontrolled cycle-to-cycle amplitude and frequency variation with more and rapidly varying turbulent air-flow. There is less discernible periodicity because of these effects and also the loss of elasticity in the vocal cords.

Where the objective voice analysis requires a voiced/unvoiced decision, this is achieved by measuring the ‘degree of periodicity’ as will be discussed later. The distinction is sometimes not easy to make especially for pathological voices for which the degree of periodicity is not very high. However many assessments of voice quality are made by analysing sustained vowels only and therefore do not require a voiced/unvoiced decision.

The requirement now is to extract and measure characteristic features of voiced sections of speech that may be indicative of abnormality, if it exists, in the operation of the vocal cords. The following features of voiced speech are commonly measured for this purpose, though there are other features that may also be of interest:

1. Fundamental frequency: This is constantly changing and characterises only pseudo-periodicity. Sometimes an average value over a period of time is sufficient, but it is sometimes necessary to have an instantaneous value of fundamental frequency which applies at a particular instant of time.

2. Shimmer: This is uncontrolled amplitude perturbation of the vocal emission. There are various different definitions of shimmer in the research literature. ‘Uncontrolled’ distinguishes the amplitude modulation from the much slower amplitude modulation that occurs naturally, for example at the beginnings and ends of words and sentences [BO00].

3. Jitter: This is uncontrolled fundamental frequency perturbation which also has

various different definitions in the literature. ‘Uncontrolled’ distinguishes the frequency-modulation from the much slower frequency modulation that occurs in natural intonation, for example in questions and tonal languages [VS09].

4. Noise-based features: These features give the perception that a random signal, often referred to as ‘noise’ or a ‘noise-like’ signal, has been added to a pseudo-periodic signal which may be already affected by shimmer and/or jitter. Such random signals can be created by turbulent air-flow via the vocal cords or elsewhere within the vocal tract. Such noise based features may be measured and quantified by the following parameters, and others:

   (a) Harmonic to Noise Ratio (HNR)

   (b) Glottal Excitation to Noise Ratio (GENR)

   (c) Normalised Noise Energy (NNE)

   These parameters measure the useful (harmonic) proportion of voiced speech (HNR) or vocal cord activity (GENR) when either or both are affected by noise due to turbulent air-flow within the larynx or elsewhere. Turbulent flow within the larynx will be amplitude modulated by the vibrations of the vocal cords.

5. Spectral features: These are features that reflect the distribution of speech energy in the frequency-domain as may be measured by a ‘short-term’ Fourier transform. Such features may be measures and quantified by:

   (a) spectral tilt

   (b) ‘low to high spectral ratio’.

Software tools for measuring features such as those mentioned above from digitised acoustic speech waveforms are available in published and commercial form for objective voice quality assessment [Kay96, Pa07, Dr99]. We now described some of these software tools. The most well known are known as ‘Praat’ [Pa07] , ‘MDVP’ [Kay96] and ‘ADSV’ [Kay96, AR06] though there are very many others which claim to do similar things and have been compared [MGS$^+$12, ULTC12, GLORSL$^+$08] with the three tools just mentioned.

2.6 ‘Praat’ software package

Praat (the Dutch word for ‘talk’) is a free software package, developed by Paul Boerma and David Wenink in the institute of phonetic science, University of Amsterdam [Pa07]. It performs speech analysis covering a wide range of standard and non-standard procedures. The package contains a number of useful measurement tools for objective voice quality evaluation. For instance, features such voiced/unvoiced decision, fundamental frequency, jitter, shimmer and HNR can be measured in a number of different ways. These effects are normally accompanied by breathiness and the perception of an additive noise-like signal within the speech. Praat measures fundamental frequency in the range of 75-600 Hz but for pathological voices the range can be extended to lower and higher values. There are various jitter and shimmer measurements provided by Praat that will now be described.

2.6.1 ‘Praat’ measurement of fundamental frequency

Praat measures the fundamental frequency (F_0) by time-locating the autocorrelation function peak that most likely corresponds to F_0. Only positive peaks are considered. The normalised autocorrelation function may be computed over a fixed time-frame, or by a normalised ‘cross-correlation’ technique as will be described in Chapter 4. Locating the correct peak can be difficult for impaired speech and errors lead to pitch doubling or halving. The Praat software goes to great trouble to try to eliminate such errors in the widest range of circumstances. Once the time-location of the correct peak has been found, its height gives an indication of the ‘degree of aperiodicity’. A height very close to 1 indicates very low aperiodicity (strong periodicity) whereas a peak height close to zero indicates the absence of periodicity. If this algorithm is applied to connected speech rather than purely voiced speech, the height, between 0 and +1, can be compared to a ‘voicing threshold’ of say 0.5, to decide whether the speech is likely to be voiced or unvoiced. This provides a voiced/unvoiced decision as referred to in Section 2.5.

2.6.2 ‘Praat’ measurements of Jitter

Equation (2.1) defines ‘absolute local jitter’ over a number, N, of pitch-cycles. It is the average absolute difference, in seconds, between consecutive pitch-periods that occurs over those N pitch-cycles. In this definition, T_i denotes the i^{th} pitch-period, in seconds, and N is the number of pitch-cycles over which jitter is to be measured.

$$Absolute\ local\ jitter(seconds) = \frac{\sum_{i=2}^{N} |T_{\mathrm{i}} - T_{\mathrm{i\text{-}1}}|}{N-1} \tag{2.1}$$

The MDVP software package (see later) calls this parameter Jita, and gives 83.200 μs as a threshold for pathology.

Equation (2.2) which we may call 'relative local jitter' over N pitch cycles. It is the average absolute difference between consecutive periods, divided by the average period. It is expressed as a percentage which is zero when all pitch periods are equal and 100% when the differences are all of the order of one pitch-period. The MDVP package calls this parameter Jitt, and gives 1.040% as a threshold for pathology.

$$Relative\ local\ jitter(\%) = \frac{100 \times N \sum_{i=2}^{N} |T_{\mathrm{i}} - T_{\mathrm{i\text{-}1}}|}{(N-1) \sum_{i=1}^{N} |T_{\mathrm{i}}|} \tag{2.2}$$

Both these measurements of jitter require reliable estimates of the fundamental periods of a succession of pitch-cycles which may not be easy to obtain for pathological voices. Relative local jitter is less sensitive than absolute local jitter to fundamental frequency estimation errors such as pitch doubling or halving as can easily occur with not very well defined periodicity. Replacing $\mathrm{T_i}$ in Equation (2.2) consistently by $2 \times Ti$ or Ti/2 does not change the relative local jitter provided it may be assumed to remain constant over the duration of the analysis. Jitter affects the frequency of the vocal tract excitation but not the resonances of the vocal tract. Therefore, looking for similarities in the shapes of consecutive cycles is a good way of detecting consecutive pitch-period differences. The Praat software gives us four other definitions of relative jitter, all of which are based on the computation of consecutive pitch-periods by this waveform-matching procedure.

A difficulty lies with the precise definition of jitter, and it is clear that any frequency-modulation will affect Equations (2.1) and (2.2) regardless of whether it is slow or rapid. A monotonic increase in pitch-period from say from 446 to 491 samples over N=10 cycles (fundamental frequency falls from 99Hz to 90 Hz over about 100 ms with Fs=44100 Hz) will produce a relative local jitter value of around 1% which is close to MDVP's threshold for pathology. A random variation of pitch-period (or an alternating sign variation) producing roughly the same magnitude cycle-to-cycle differences produces a very similar value of relative local jitter. The relatively slow monotonic increase could occur in normal speech, for example at the end of a statement or during a Chinese word. A corresponding increase in frequency might occur at the end of a question. Hence these definitions of jitter may classify natural intonation patterns as jitter and indicative of pathology. The random or alternating variations in pitch-period are much faster and are more likely to be indicative of abnormal voice, and therefore classifiable as jitter. Although no precise definition of jitter, in terms of the nature and speed of the frequency modulation, have been found in the literature, the problem referred to here has been noticed and remedied to a degree.

Equation (2.3) defines a variation of relative local jitter which is known as Relative Average Perturbation (RAP). It is the average absolute difference between a period and the average of it and its two neighbours divided by the average period. The result is normally represented as a percentage which is zero when all pitch-periods are equal and 100% when differences are of the order of the pitch-periods.

$$Jitter(RAP) = \frac{\sum\limits_{i=2}^{N-1} |T_{\mathrm{i}} - (T_{\mathrm{i\text{-}1}} + T_{\mathrm{i}} + T_{\mathrm{i+1}})/3|/(N-2)}{\sum_{i=1}^{N} T_{\mathrm{i}}/N} \tag{2.3}$$

A further variation of relative local jitter is called Five-Point Period Perturbation Quotient (PPQ5) which is defined by Equation (2.4). The absolute differences between a period and the average of it and its four closest neighbours is computed and divided by the average period. The result is normally expressed as a percentage.

$$Jitter(PPQ5) = \frac{\sum\limits_{i=3}^{N-2} |T_{\mathrm{i}} - (T_{\mathrm{i\text{-}2}} + T_{\mathrm{i\text{-}1}} + T_{\mathrm{i}} + T_{\mathrm{i+1}} + T_{\mathrm{i+2}})/5|/(N-4)}{\sum\limits_{i=1}^{N} T_{\mathrm{i}}/N} \tag{2.4}$$

The jitter measurements analysed above are well known as a sort of standard and for this reason they will be adopted in this thesis.

DDP is the average absolute differences between consecutive periods. Equation (2.5) the definition of DDP which is normally expressed as a percentage.

$$Jitter(DDP) = \frac{\sum\limits_{i=2}^{N-1} |(T_{\mathrm{i+1}} - T_{\mathrm{i}}) - (T_{\mathrm{i}} - T_{\mathrm{i\text{-}1}})|/(N-2)}{\sum\limits_{i=1}^{N} T_{\mathrm{i}}/N} \tag{2.5}$$

Clearly larger values on N may be used when analysing sustained vowels than would be appropriate for connected speech. Although much of the work in this basis is based on the analysis of sustained vowels, we wish to make the methods devised also appropriate for analysing the voiced sections of connected speech that can be identified by voiced/unvoiced detection methods. The reduced sensitivity of the RAP and PPQ5 estimates to slow pitch changes that occur in natural intonation is clearly more important for connected speech than for sustained vowels. However fundamental frequency changes do occur in sustained vowels as well. It is very hard to maintain a fixed frequency for any length of time, and there is likely to be 'wavering', 'tremor' or 'vibrato' in the voice which is natural and not indicative of any voice problem. Tremor is defined as non-monotonic fundamental frequency modulation which is slower than jitter and does not create roughness. Vibrato is similar, but perhaps more controlled. RAP and PPQ5 estimates should successfully de-emphasize the effect of tremor and vibrato on their estimates of jitter in favour of uncontrolled higher frequency and more random changes.

2.6.2.1 'Praat' measurements of shimmer

Relative local shimmer, as defined by Equation (2.6) over N pitch-cycles, is the average absolute difference between the amplitudes A_{i} of consecutive pitch-cycles divided by the average pitch-cycle amplitude [Pa07]. The amplitudes are the maximum values within the cycle and assumed to be proportional to the root mean square value of the corresponding vocal tract excitation cycles.

$$Shim = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_{\mathrm{i}} - A_{\mathrm{i+1}}|}{\frac{1}{N}\sum_{i=1}^{N}A_{\mathrm{i}}} \tag{2.6}$$

Absolute local shimmer is generally defined in terms of a decibel representation of the amplitudes rather than absolute amplitudes. Equation (2.7) defines ShdB (a decibel form of shimmer) as the variability of the amplitudes in dBs averaged over N pitch-cycles [Pa07].

$$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1}|20 \times \log\frac{A_{\mathrm{i+1}}}{A_{\mathrm{i}}}| \tag{2.7}$$

The Three-Point Amplitude Perturbation Quotient (APQ3) defined by Equation (2.8) is the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its neighbour divided by the average amplitude.

$$Shimmer(apq3) = \frac{(\frac{1}{N-2})\sum_{i=2}^{N-1}|A_{\mathrm{i}} - (A_{\mathrm{i}} + A_{\mathrm{i\text{-}1}} + A_{\mathrm{i+1}}/3)|}{\frac{1}{N}\sum_{i=1}^{N} A_{\mathrm{i}}} \tag{2.8}$$

The Five-Point Amplitude Perturbation Quotient (APQ5) given by Equation (2.9) describes average absolute difference between the amplitude of a period and the average of the amplitudes of and its neighbour divided by the average amplitude

$$Shimmer(apq5) = \frac{(\frac{1}{N-4})\sum_{i=3}^{N-2}|A_{\mathrm{i}} - (A_{\mathrm{i}} + A_{\mathrm{i\text{-}2}} + A_{\mathrm{i\text{-}1}} + A_{\mathrm{i+1}} + A_{\mathrm{i+2}}/5)|}{\frac{1}{N}\sum_{i=1}^{N} A_{\mathrm{i}}} \tag{2.9}$$

The Equations used for shimmer use essentially the same smoothing techniques as were used for jitter, and for the same reason. Slow changes in amplitude as occur naturally, for example at the beginnings and ends of words and sentences are de-emphasised in favour of less controlled rapid changes of amplitude.

2.6.3 Praat measurements of noise-based features

Praat defines the 'harmonicity' as a degree of acoustic periodicity and noise-based features. This also called Harmonics-to-Noise Ratio (HNR) and it is expressed in dB. For examples if 98% of the energy of the signal is in the periodic part, and 2% is noise, the HNR is $10 \times log10(98/2) = 17dB$. The equal energy in the harmonics and in the noise makes HNR equla to '0'. Praat Uses Autocorrelation [Pa07] and Cross correlation method for measuring the 'harmonicity'.

2.7 Multi-Dimensional Voice Program (MDVP)

A software package known as the 'multi-dimensional voice program' (MDVP) has been developed and commercialised [Kay96] for the analysis of both digitised acoustic voice and EGG waveforms. This software is commercially available from two main sources: Laryngograph and KayPentax [Kay96] and provides many measures of different aspects of voice quality. Some measurements are presented in the form of numerical data, but many are produced in graphical form. MDVP calculates measurements of more than 25 voice features for each single vocalisation. The large number of features is provided on the grounds that no single feature can be universally appropriate. Unfortunately, the multiplicity of features does not provide an obvious way of determining an overall equivalent to the GRBAS assessment. Each of the multiple features represent a very specific facet of voice quality with no overall universally accepted measurements as is provided by GRBAS. Some of the features measured by MDVP are in the following sections.

2.7.1 'MDVP' measurements of fundamental frequency

Although the precise details are not published, MDVP computes F_0 by much the same procedure as the Praat software. Values of F_0 obtained are reported to be comparable over a wide range of voices, though voiced/unvoiced decisions are reported to be significantly different [MGS+12] because of different voicing thresholds.

2.7.2 'MDVP' measurements of jitter

Absolute jitter (jita) is an evaluation of the cycle-to-cycle variability of the pitch-period within the analysed voice sample. The MDVP definition of jita is widely in the literature [SOA09, CTPB+00], and is the same as the Praat definition of 'absolute local jitter' given in Equation (2.1). Its units are in seconds. Relative jitter (Jitt) is the same as relative local jitter defined by Praat and given in equation (2.2). Relative Average Perturbation (RAP) is the same as the RAP version of jitter defined by Praat and presented in Equation (2.3). The 'Pitch Period Perturbation Quotient' (PPQ) quantifies period-to-period variability with a smoothing factor of 5 periods as with the Praat PPQ5 measure defined by Equation (2.4). The higher smoothing factor leaves PPQ less sensitive to natural period-to-period variations [MDV]. The Smoothed Pitch Period Perturbation Quotient (SPPQ) defined by Equation (2.10) is a generalisation of RAP and PPQ5 where the order of the smoothing process may be defined by the user. The factory setup for the smoothing process is 55 periods but this may be changed. Voice break areas are automatically excluded.

$$SPPQ = \frac{\frac{1}{N-sf+1}\sum\limits_{i=1}^{N-sf+1}\left|\frac{1}{sf}\sum\limits_{r=0}^{sf-1}{T_{\mathrm{O}}}^{(i+r)} - {T_{\mathrm{O}}}^{(i+m)}\right|}{\frac{1}{N}\sum\limits_{i=1}^{N}{T_{\mathrm{O}}}^{(i)}} \tag{2.10}$$

The PPQ measurement of jitter is considered to describe well the cycle-to-cycle irregularity associated with the inability of the vocal folds to support a periodic vibration with a defined period. Hoarse and/or breathy voices will have an increased PPQ. MDVP provides the jitter estimates ‘jita’, ‘jitt’ and ‘Jita’ in addition to PPQ because the research literature contains normative data for these measurements.

Both MDVP and Praat software packages are capable of producing estimates of the amount of jitter in a sustained vowel. Some of the MDVP algorithms have a tendency higher values of jitter than the Praat algorithms. When applied to the same speech segment they provide different estimates [MGS+12, MCDB+09, HKŞ11, BKG+96]. Apart from the methods in these two software packages there are other methods for estimating jitter [KK90, VMJ96]. The question is how to compare them.

2.7.3 ‘MDVP’ measurements of shimmer

The MDVP software package also provides measurements of shimmer which correspond broadly to the methodology used in Praat and outlined earlier, but with higher degrees of smoothing recommended. The Amplitude Perturbation Quotient (APQ) is a relative evaluation of the cycle-to-cycle variability of the peak amplitude within each cycle. It is defined by equation (2.11) where Ai, for i=1,2,...N are the extracted amplitudes and N is the number of pitch-cycles. It uses 11-cycle smoothing expressed in a slightly different way.

$$APQ = \frac{\frac{1}{N-10}\sum\limits_{i=1}^{N-10}\left|\frac{1}{11}\sum\limits_{r=0}^{10}A^{(i+r)} - A^{(i+5)}\right|}{\frac{1}{N}\sum\limits_{i=1}^{N}A^{(i)}} \tag{2.11}$$

The Smoothed Amplitude Perturbation Quotient (SAPQ) is similar to equation (2.11) but with a degree of smoothing that may be defined by the MDVP user. Using SAPQ the MDVP user can compare his amplitude perturbation results with other results in the literature such as are obtained using different Praat measurement. The smoothing factor determines the number of cycles used for the smoothing and may be selected in the range from 1 to 199 A general formula for SAPQ is given in Equation (2.12) where sf is the smoothing factor.

$$SAPQ = \frac{\frac{1}{N-sf+1}\sum_{i=1}^{N-sf+1}|\frac{1}{sf}\sum_{r=0}^{sf-1}A^{(i+r)} - A^{(i+m)}|}{\frac{1}{N}\sum_{i=1}^{N}A^{(i)}} \tag{2.12}$$

Instability of the fundamental frequency F_0 and amplitude instability tend to increase with age voice, resulting in greater jitter and shimmer values, tremor and increased hoarseness [LCB82]. Few studies have investigated F_0 effects, Baken and Orlikoff [OB90] concluded that the influence of F_0 on jitter and shimmer has not been fully understood to date.

2.7.4 'MDVP' measurements of noise-based features

MDVP defines two features for measuring noise in voice signals. These are Voice Turbulence Index (VTI) and Noise-to-Harmonic Ratio (NHR). They may be described as follows.

2.7.4.1 MDVP measurements of Voice Turbulence Index

Voice Turbulence Index measures noise turbulence caused by incomplete closure of the vocal folds. Pitch synchronous frequency-domain methods are used for the extraction of VTI.

1. MDVP computes an unwindowed 1024-point Fast Fourier Transform (FFT) for the data. This is converted to a power spectrum.

2. The fundamental pitch-frequency is calculated using a synchronous pitch extraction method.

3. MDVP separates the spectrum into the harmonic and inharmonic components synchronously using the average fundamental frequency. This is computed for 1024 sample segments of speech.

4. The VTI a speech segment is the ratio of the inharmonic spectral energy in the range 1800-5800 Hz to the harmonic energy in the frequency range 70-4200 Hz.

2.7.4.2 'MDVP' measurements of noise-to-harmonic ratio (HNR)

MDVP defines the Noise-to-Harmonic Ratio (NHR) of a section of voiced speech as the average ratio of the non-harmonic spectral energy to the harmonic spectral energy in the frequency range 70 Hz to 4200 Hz. In other words, NHR measures the relative contributions of aperiodic and periodic components of the voice signal. Normal periodic voiced signals have a small NHR whereas severely dysphonic voiced signals that have high components of breathiness and roughness tend to have higher values of NHR. Harmonic-to-noise ratio (HNR) is the reciprocal of NHR. It is claimed [MDV] that a pitch-synchronous frequency-domain method is used by MDVP for NHR computation, though precise details are not published. The use of pitch-synchronous Fourier analysis eliminates the spectral spreading that normally occurs with fixed analysis block-lengths. A viable pitch-synchronous procedure that is likely to be similar, in principle, to that used by MDVP proceeds as follows:

Firstly, the voiced speech segment is divided into frames each containing approximately 20 ms of speech. Then the fundamental pitch-frequency is calculated for each frame using a highly reliable F_0 extraction method. An FFT with rectangular window is then applied to an integer number of complete pitch-cycles. The number of complete cycles could be as low as two, but there may be advantages in taking a few more when the pitch-frequency is not changing rapidly. The more pitch-cycles taken, the more accurate the result, potentially, but the more susceptible will be the analysis to the effects of aperiodicity and non-stationarity. The FFT bock-length must therefore vary as the pitch-frequency changes, and an accurate estimate of the pitch-frequency is essential to make this approach work. The pitch-synchronous block-length must be chosen such that the characteristics of the speech within it remain close to stationary. When this is the case, the FFT magnitude spectrum will have lines at the fundamental frequency and its harmonics, with no spectral spreading in-between. There will be nothing in-between the spectral lines when there is no noise and the speech is purely periodic. When there is noise, its power may be estimated from the magnitude spectral samples between the harmonics assuming the noise is spectrally white or of another spectral shape. Once the power of the noise has been estimated as from a standard periodogram, it may be subtracted from the overall spectral power to obtain the harmonic power and hence the noise-to-harmonic ratio or its inverse. The harmonic power, thus calculated, should be close to the sum of the harmonic powers, but not exactly equal to it because of noise added to the harmonics.

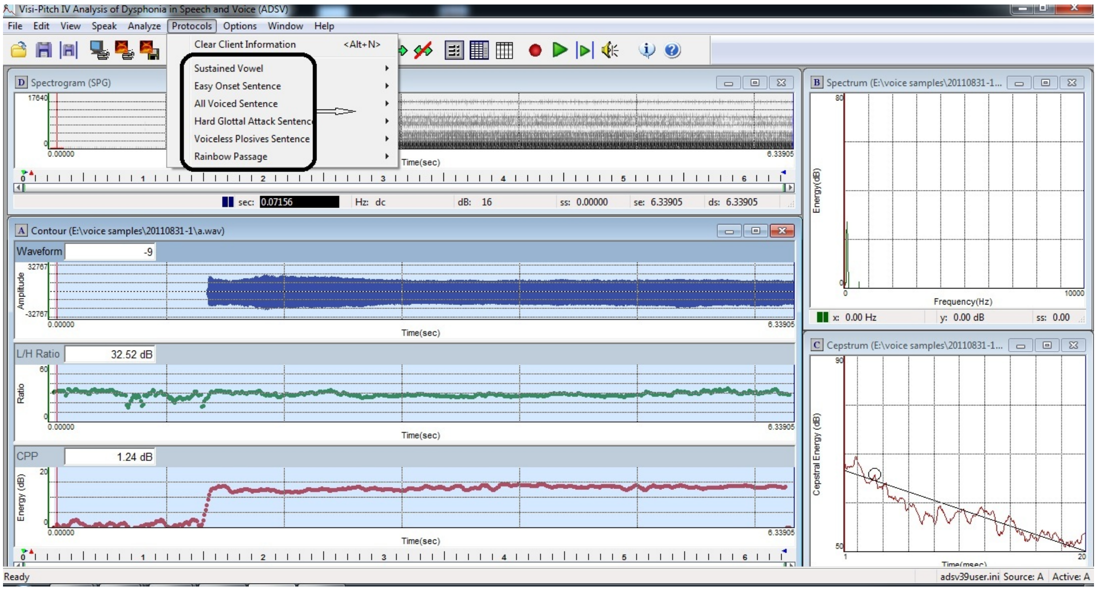
Clearly, the NHR estimate will be affected by any degree of aperiodicity, including shimmer, and jitter. There are various tricks that may be applied. For example, the number of pitch-cycles can be chosen dynamically; small during periods of transition and larger when the speech is highly stationary. Also, dynamic cycle-stretching (not 'time-warping') and amplitude scaling may be employed to reduce the effects of jitter and shimmer and make the pitch cycles appear more periodic. It is not known how many of these tricks are employed by MDVP and it what way. In principle, pitch-synchronous frequency domain techniques can be made exactly equivalent to any time-domain or autocorrelation-domain technique. Anything done in the time-domain can also be done in the frequency-domain, and any difficulties encountered in one domain will also manifest themselves, in some different form, in the other domain. It has been reported [MGS$^+$11] that the values of HNR obtained from Praat and MDVP are usually significantly different. Comparisons with other voice analysis programs have also revealed similar differences [MGS$^+$12]. A possible reason for such differences is that MDVP calculates the NHR in the frequency range 70-4200 Hz.

2.8 Analysis of Dysphonia in Speech and Voice (ADSV)

'Analysis of Dysphonia in Speech and Voice' (ADSV) is a commercial software package for objective voice analysis [AR06]. It can perform objective voice assessment on recordings of sustained vowels and continuous speech within normal and mild-to-severely dysphonic voices. ADSV uses spectral and cepstral based analyses, which overcome some problems of traditional acoustic assessment methods in Praat and MDVP that are dependent on identifying individual pitch cycles. ADSV can provide valid and reliable voice quality assessments of non-periodic voice segments as they occur in samples of continuous speech (sentences) and severely dysphonic speech [AR09, AR06, AR05]. Graphical displays showing how spectral and cepstral values change over time are provided.

ADSV provides protocols for measuring samples of speech with particular characteristics. The protocols in ADSV require samples of sustained vowels, 'easy onset' sentences, 'all voiced' sentences, 'hard glottal attack' sentences, 'voiceless plosive' sentences and a 'rainbow passage' consisting of a standard piece of text. Figure 2.8 shows a screen-shot of part of the ADSV graphical user interface and it lists the six protocols that can be used in ADSV [ADS]. The protocols are as follows.

1. Sustained vowel: The aim of this protocol is to assess the ability of the participants to produce sustained and effective voicing in a steady pitch and loudness


Visi-Pitch IV Analysis of Dysphonia in Speech and Voice (ADSV)
File Edit View Speak Analyze Protocols Options Window Help
Clear Client Information <Alt+N>
Sustained Vowel
Easy Onset Sentence
All Voiced Sentence
Hard Glottal Attack Sentenc
Voiceless Plosives Sentence
Rainbow Passage
D Spectrogram (SPG)
17640
0
0.00000
6.33905
Time(sec)
0 1 2 3 4 5 6
B Spectrum (E:\voice samples\20110831-1...
80
Energy(dB)
0
10000
Frequency(Hz)
A Contour (E:\voice samples\20110831-1\a.wav)
Waveform
Amplitude
32767
-32767
L/H Ratio
Ratio
60
CPP
Energy (dB)
20
C Cepstrum (E:\voice samples\20110831-1...
Cepstral Energy (dB)
90
50
1
20
Time(msec)
Ready

sec: 0.07156 Hz: dc dB: 16 ss: 0.00000 se: 6.33905 ds: 6.33905

x: 0.00 Hz y: 0.00 dB ss: 0.00

Waveform -9

L/H Ratio 32.52 dB

CPP 1.24 dB

adsv39user.ini Source: A Active: A

Figure 2.8: This is a windowing containing ADSV data. The green data mark at the start of the data. The red data cursor has been moved to the instance of L/H Ratio and CPP data located. This is reported in the x-axis box located on the status line. The waveform, L/H Ratio contour, and CPP contour are displayed. ADSV Results are calculated from the blue highlighted data located between the selection cursors.

context. (e.g. /a/ or /e/)

2. Easy Onset Sentence: The aim of this protocol is to elicit voice characteristics such as soft glottal attacks and voiceless to voiced transitions. (e.g.‘How hard did he hit him?’)

3. All Voiced Sentence: The aim of this protocol assess the presence of possible spasms or voiced stoppages and the ability to maintain consistent voicing during connected speech. (e.g.‘We were away a year ago’).

4. Hard Glottal Attack Sentence: The purpose of this protocol is to assess the presence of hard glottal attacks. (e.g.‘We eat eggs every Easter’)

5. Voiceless Plosives Sentence: The aim of this protocol is to assess the ability to transition easily between vowel production and voiceless stop-plosive production. (e.g. ‘Peter will keep at the peak’).

6. Rainbow Passage: The purpose of this protocol is to evaluate the voicing capability in a obtained from a traditionally used phonetically-balanced passage.

2.9 Discussion

The three software packages mentioned in this Chapter, i.e. Praat, MDVP and ADSV, use various methods for measuring voice features. Both Praat and MDVP use an autocorrelation method for pitch analysis, Praat uses the original amplitude for pitch analysis whereas MDVP quantities the amplitude into the values -1, 0, +1 before computing the autocorrelation [Pa07]. Therefore, voice features analysis can produce different results when using these programs. Eventually, they do not provide an obvious way of determining an overall equivalent to the GRBAS scale because of the multiplicity of the features.

2.10 Other software packages for speech analysis

WPCVox is a commercial tool for recording and analysing speech and electroglottographic(EGG) signals [GLORSL$^+$08]. WPCVox permits the synchronous recording of speech and EGG signals using an active connector . It also mixes both signals together. This represents graphically the features for voice quality assessment. A Paper by [GLORSL$^+$08] concluded that the results obtained for WPCVox very similar to those obtained with MDVP.

Dr.Speech (DRS) [Dr99] is a commercial tool for analysing speech. Dr. Speech is rarely used in the literature but smaller voice clinics and students often use it [SCDB05]. A paper by [SCDB05] conclude that both Dr.Speech and MDVP generate comparable results for F_0, shimmer, and HNR for normal adult. For normal adult voices, F_0SD, absolute, and relative jitter, the results of both programs are not comparable [SCDB05].

2.11 Comparisons between Praat, MDVP and other packages

Paper [HKŞ11] compares the acoustic analysis results obtained by the Praat and MDVP software suites for a selection of 47 voice samples consisting of both normal and pathological voices. Each of the two software suites was used to obtain mean fundamental frequency, jitter, shimmer, and harmonics to noise ratio. The measurements obtained for mean fundamental frequency and shimmer were not significantly different. However, measurements for jitter and noise-to-harmonics ratio were significantly different. The paper reports strong correlation between the Praat and MDVP measurements of jitter (according to all four definitions: (abs, local, RAP, PPQ) despite the numerical differences. MDVP seems to consistently give higher values of jitter than Praat. The agreement and correlations for shimmer values (dB, local and apq) were found to be moderate. Although the Praat and MDVP measurements of harmonics-to-noise ratio (HNR) are very different, the correlation was moderate. It is presumed, though not stated in the paper, that the measure of correlation is Pearsons which, as will be discussed in Chapter 3, reflects trends from the individual means rather than absolute differences. These findings, and similar results from other authors, present difficulties for the work in this thesis if it is to make use of measurements obtained by published and commercially available software suites such as Praat and MDVP.

The findings for jitter are surprising in view of the fact that both programs use the same definitions for the four versions (abs, local, RAP, PPQ). The paper [HKŞ11] believes that the reason for the different results may be the different voiced cycle detection algorithms used by Praat and MDVP [BW04, MDV]. Even this comment is surprising since the measurements of fundamental frequency are comparable. Nevertheless, the moderate correlation between the numerically different parameters appears to indicate that both computer programs use similar strategies for normal and pathologic voices. Unfortunately, it is difficult to analyse these differences further since the details of the algorithms used are not published or clearly documented for either software suite.

The differences between the Praat and MDVP measurements of jitter are also highlighted in a paper by Maryn et al [MCDB$^+$09] which concludes that the two programs give different results for both jitter and shimmer. It is reported that MDVP gave consistently higher measures than Praat for the four different measures of jitter (absolute relative, RAP and PPQ) and three measures of shimmer (dB, relative and apq) for 50 subjects with various voice disorders.

In the same journal issue, a paper by Paul Boersma [Boe09] attempts to explain which of the two software suites, Praat or MDVP, gives the best result. Paul Boersma is one of the inventors of Praat [BW04]. However, the paper [Boe09] sets out to justify its claim that Praat is more reliable than MDVP for computing jitter by explaining why the suspicions of Maryn et al [MCDB$^+$09] are justified. These suspicions were that the methods used for locating the time-locations of the vocal tract excitation pulses were the source of the differences.

The standard method used by Praat is referred to as ‘wave-form matching’ whereas the MDVP method is ‘peak picking’. Waveform matching identifies complete individual pitch-cycles by their entire shapes between the time-locations of successive excitations. There is similarity between the shapes of complete pitch-cycles because of the resonances of the vocal tract. These resonances, being controlled by the physical shape of the human vocal tract, can only change relatively slowly in comparison to the waveform changes that occur due to noise, jitter and shimmer. Therefore the similarity of successive pitch cycles can be exploited to accurately determine the duration and amplitude of each pitch-cycle which is required for calculating jitter and shimmer.

Peak picking is rather simpler than waveform matching and simply looks for the time-locations where the speech waveform has its local maxima. The time-location of each maximum is assumed to correspond exactly with the snapping closed of the vocal cords. The amplitudes of successive maxima are used to calculate shimmer. The inevitable slight delay between the ‘snapping closed’ and the peak is not important.

The shape of each pitch-cycle, and also the sharpness of the peaks, will change marginally with the effect of jitter and shimmer, because of the slightly different interaction of each new pitch-cycle with the still decaying resonances of previous cycles. Pitch-cycles do not die away completely at the end of the cycle and ‘ring on’ into subsequent cycles. But this effect is usually not significant. Therefore average pitch-period estimates and values of jitter and shimmer are usually comparable between Praat and MDVP when the degree of jitter or shimmer is not excessive, i.e. less than about 10%, and when there is no noise-like’ signal added at the vocal cords or further along the vocal tract. This result also requires that there is no background noise in the recording.

However, according to Paul Boersma [Boe09], the presence of even small amounts of noise, (e.g. 1%) due to turbulence at the vocal cords (breathiness) or elsewhere creates problems for the peak-picking approach which are much less serious for the waveform matching approach. Essentially the noise creates uncertainty in the exact time-locations and amplitudes of the peaks which are critical for peak picking approaches. In contrast, because the time-duration and amplitude of each pitch-cycle is estimated from the whole cycle and not just one peak value within it, the effect of the added uncertainty tends to be averaged out. This averaging out is effective when using the difference between successive pitch cycles to estimate the pitch-period and amplitude of each cycle. Note that only the pitch-period is important for estimating shimmer and jitter; the exact time-locations of the starts and ends of each pitch-cycle are not important.

Paul Boersma [Boe09] presents results which show that for speech sounds with moderate jitter and turbulence or other random effects adding 1% to the overall power, the differences in jitter estimates are significant between waveform matching and peak picking. For example, simulated jitter at 0.09 % is estimated as 0.076 % jitter by waveform matching (Praat) and 0.518 % by peak-picking (MDVP).

2.12 Use of existing software packages in this thesis

The findings of paper [HKŞ11] and others highlight a difficulty of using published or commercial voice analysis suites such as Praat and MDVP for deriving the measurements of features required for the objective analysis of voice. Although, this software is highly sophisticated and, doubtless, endlessly evaluated and optimised, it is not possible to use this software in this thesis without an understanding of the underlying DSP algorithms they employ. Indeed, the Praat software suite offers choices between different techniques for the same measurements, with only a cursory explanation of the essential differences. The elaborate nature of the software also makes it very difficult to examine minimally documented source code to extract and examine the required techniques. The fact that the software suites produce different results makes the choice between them a difficult decision.

Further practical difficulties also emerge with the desire to produce, in this thesis, the prototype of an integrated software package for objective voice quality analysis. If such a package were to obtain its feature measurements from Praat or MDVP it would need a convenient interface which is not currently available. Both suites are driven from graphical user interfaces which require user intervention. Praat does have a scripting option which, in principle, allows data from a different software package to be analysed. But current versions (e.g. Praat version 5.4.19) does not allow all the necessary operations to be controlled from a script. There are other voice analysis suites such as Dr. Speech or Vox, but they have been reported as suffering from the same problems as the Praat and MDVP software.

The techniques required for extracting and measuring features that are likely to reflect speech quality are generally well known. They employ basic DSP operations such as digital filtering, Fourier analysis and forms of autocorrelation analysis. The cepstrum, which is derived from Fourier analysis, is widely used. Details of these approaches will be given in Chapter 4, along with a description of prototype software that has been developed to implement viable versions of the feature measurement techniques needed in this thesis. This prototype software uses, we believe, the best available approaches. However, a full optimisation up to the level of commercial software has not been done.

Most sensible speech feature measurement techniques, even simple ones, will work for most normal voices and some moderately pathological voices. Even very sophisticated techniques will fail for very severely damaged or abnormal voices. The challenge is to devise techniques that work well for the widest possible range of voices. Selecting and implementing the best DSP approach is only the first step in meeting this challenge. The next stage is painstaking optimisation over the very wide range of possible voice impairments, finding the causes of failure and trying to eliminate these causes.

2.13 Machine learning algorithms

2.13.1 Type of learning

Objective voice quality analysis would be useful in many clinical applications. Such methods are already available in the form of Praat, MDVP and ADSV software packages, but the results they produce are not easily understood by clinicians. Therefore a technique that can map the results of objective analysis techniques to the classifications produced by more familiar subjective methods, such as GRBAS, is needed. Looking at the machine learning area for such mapping techniques should be useful.

In machine learning, supervised and unsupervised learning methods distinguish the two major learning models. Supervised learning methods attempt to discover the relationships between input features and target attributes [KZP07]. For each observation of the predictor measurements, for example the voice measurements made by the Praat, MDVP and ADSV software, there are associated response measurements, such as GRBAS scores. The aim is to fit a model that relates the responses to the predictors, with the aim of accurately predicting the responses for future observations (prediction) or better understanding the relationship between the responses and the predictors (inference). Unsupervised learning methods are methods for which for every observation produces a vector of measurements but without labeled responses for the target [HTF09]. In this thesis, we focus on supervised learning methods.

The two main models used for mapping the observations to the target outputs (such as GRBAS scores) by supervised learning methods are ‘Classification’ and ‘Regression’. Classification maps the observed ‘features’ into pre-defined classes whereas regression models map the features on to a real-valued numerical domain. There are many classification and regression models such as K-Nearest-Neighbor (KNN), decision trees, neural networks and support vector machines for representing classifiers. Multiple linear regression and logistic regression are commonly used regression models, and KNN can be used for either classification or regression.

2.13.2 How supervised learning algorithms work

A supervised learning algorithm receives a set of examples of pre-analysed data for which the required target properties are known, and then tries to learn how predict the corresponding target properties of other data, which are unknown. The original set of examples provides training data. Once the learning algorithm has been trained using all or some of this data, if the same data is re-used to compute the error of the model fit, an overly optimistic estimate of the error of the model will be obtained. This is because the training or model fitting process tries to ensure that the error of the model for the training data is as low as possible.

Therefore, the model will be specifically suited to the training data. To get a more realistic estimate of how the model will perform with unseen data, some unseen data which is not used for the training process must be available along with its known properties for checking. It is therefore common practice to set aside part of the original data for checking and not to use it in the training process. This data-set is may be called the validation data-set. After training the model using the remaining data-set, now known as the training data-set, the performance may tested or ‘validated’ on the validation data-set [VV98, KZP07].

The validation data-set is often used to fine-tune models. For example, we may try out various sets of coefficients for a regression model by finding the error produced by each set of coefficients for the validation data-set. This would allow us to choose among the competing sets of coefficients. In such a case, the error with the validation data-set will be an optimistic estimate of how the fine-tuned model would perform with unseen data. This is because the final coefficients will have been chosen such that the error with the validation data-set is the lowest possible. Thus, we may need to set aside yet another portion of the original data which is used neither in training nor in validation. This set may be called the test data-set to distinguish it from the validation data-set. The error produced by the fine-tuned model applied to the test data then gives a realistic estimate of the performance of the model on completely unseen data.

2.13.3 Cross-Validation

There are many ways of choosing the preferred sizes of training, validation and testing data-sets. A well accepted method is N-Fold cross-validation, in which the order of the original data is randomised and then N equal size partitions are made [K^+95]. The data is split into several parts which are called folds. For example, for 5-fold cross-validation, in the first step, a model may be trained and possibly fine-tuned on folds 1-4 and then tested on fold 5. To improve the model without the requirement of further data-sets, a second step may be performed whereby the model is retrained and tuned on folds 1, 2 and 3 and 5, and then retested on fold 4. This process may continue by re-training and tuning is on folds 1, 2, 4, 5, then testing on fold 3 and so on. In each step, the fold that is left out is not seen by the model until the testing phase. In each fold, the error will be averaged and called the cross-validation error. This cross validation error will be an optimistic estimate of what may be expected with truly unknown data because the parameters will implicitly be fitted to the training data by iterating the training process over the same folds. A better solution may be obtainable by separating the data into three different parts as outlined earlier: a training set, a validation set and test set. Therefore, the recommended procedure for training a GRBAS predictor from a data-set containing 102 previously scored examples is as follows.

1. The 102 examples are expressed as a matrix where the rows correspond to the subjects or patients and the columns contain the measurements of voice features (the observations) and the GRBAS scores produced by trained scorers.

2. If there is more than one scorer (actually we had five scorers) a consensus or 'gold standard' (see Chapter 3) score must be agreed for each GRBAS component for each subject by some form of averaging.

3. After randomising their order, the rows are split into N folds where N, for example, may be equal to 10.

4. The last fold is kept as a 'hold out' test set.

5. The cross-validation is performed with the remaining 9 folds.

6. The model that performs best on average over those 9 folds is selected.

7. The model is then applied to the ‘hold-out’ test set for evaluation. The error on the hold-out data is an unbiased estimate for the future generalisation-error.

We can repeat the procedure several times to remove the variance of the whole the procedure. In Chapter 5, the experiment for training and testing will be repeated for 20 times (trials).

2.14 Objective voice quality analysis conforming to the GRBAS scale

The accurate assessment of pathological voice quality is a major research problem that has attracted attention in the field of voice disorder and biomedical engineering for many years. Voice quality assessment using subjective methods are based on a trained listener’s opinion of the quality of an utterance. There will normally be just one listener for a patient, and the patient must rely on the professionalism and expertise of the listener for an appropriate assessment or score. It must be assumed that listeners are self-consistent in their assessments and also that different trained listeners around the country, by virtue of their training, will give similar scores for similar degrees of voice degradation. The extent to which these assumptions may be reliable is tested in this thesis since five scorers were available as part of the research effort and they were required to score a small proportion of the subjects twice.

It is possible that objective computerised assessment could replace or augment the expensive and time-consuming manual procedures currently required, and that computers could produce similar results with similar or perhaps even better consistency. Although it is not suggested that objective methods should completely replace subjective ones, the objective quality evaluations may be very useful, and there is some evidence already that the results can be made to correlate well with subjective quality assessments. There are few studies about objective voice quality assessment conforming to the GRBAS scale.

A recent paper [VCOAAL^{+}13] uses a K Nearest Neighbour classifier to predict all parameters using spectral energy measurements, cepstral coefficients, a glottal-to-noise excitation ratio and other features. The objective scores were compared with perceptual evaluations by a single expert at the University Poletecnicaof Madrid. Good correspondence was obtained, the best efficiency being obtained for Asthenia was 89.3% [VCOAAL^{+}13].

The work by [GHSS05] used three features to classify speech signals into a three-point rating scale by considering only the G component from GRBAS scale. The 10 patients with Parkinsons disease and four healthy speakers make the database. A measure called 'Itakura-Satio distortion' provides good correlation with the perceptual evaluation and could be used to predict it. No classification results were presented and the number of participants in the database was very small.

Nicolas Saenz et al [SLGLORGV06] used Learning Vector Quantization (LVQ) and a KNN classifier for predicting all GRBAS component. MFCC were extracted as features. The voice examples were scored by three ENT clinicians and they analysed a short-time EGG signal. The most accurate results were obtained for 'control' participant with 65% accuracy and for class '0' and class '1'.

Other researches in the literature use different classification techniques for the detection of pathological voices. These are as follows.

The classification of normal and pathological voices was carried out with a Multi-Layer Perceptron (MLP) neural network by [FSLGL$^+$09]. The experiments were performed using a subset of the MEEI [EI94] database with 53 normal and 173 pathological speakers and the participants were differentiated by sex. A classification accuracy of 88.3% was obtained. The feature set used to train the ANN(artificial neural network) based detector was based on MFCC measurements.

A modification of the standard KNN classifier was proposed by [SC$^+$07] to classify a set of 163 pathological and 53 normal speakers extracted from MEEI [EI94] database. The best accuracy obtained was 94.28% by using HNR in four frequency bands.

A classifier based on a least square Support Vector Machine (SVM) with three different kernel functions was used to identify laryngeal pathologies by [FGS$^+$07]. The features used to train the classifier are statistics estimated from linear prediction coefficients and time-frequency representations using wavelet decompositions. The experiments were carried out using a data set composed of 30 normal and 30 pathological participants. Classification accuracy up to 91.67% was obtained

To differentiate between normal and pathological voices a probabilistic model, called Gaussian Mixture Model (GMM) was applied by [GLGVBV06]. The obtaining an efficiency around 94% with the same data set (MEEI) used in [FSLGL$^+$09] and the features used to train the classifier were MFCC along with their first derivative.

A work by [DNB$^+$02] propose an automatic detection of 'normal' and 'pathlogical' speech from sustained vowels. MFFCC and pitch dynamic were useed as the features. GMM and Hidden Markov Model (HMM) classify speech into 'normal' and ‘pathlogical’ categories . The experiment was applied to the MEEI data-base. The best obtained accuracy was about 99.4% with extracted features

Gelzinis et al [GVB08] used 11 different sets of features, including noise measures, energy perturbation measures estimated from different frequency bands and linear predictive coefficients. They extracted 23 measures using a commercial software called ‘Dr. Speech’. The experiments were carried out using 79 pathological and 69 healthy speakers. They combine six SVM trained with different sets of features. The best achieved classification rate was about 95.5%.

An approach for the objective voice quality assessment was presented by [RMM02]. This work is based on a seven-point ranking scheme and Artificial Neural Network (ANN). From EGG signals different combinations of short-term and long-term time-domain and frequency-domain features were extracted. A database was composed of 77 pathological speech signals . The best result was obtained using 21 input features. An average accuracy of 92% was obtained.

Despite all approaches for the objective voice quality assessment found in the literature, their results can not be easily compared. Our work aims to use a different database, five experienced SLT scorers, analysing acoustic signal and a different feature set. Also, we aim to use regression models rather than classification, and to compare two regression models. Regression is sensitive to the degree of disagreement between scores where classification is concerned only with agreement or disagreement.

Tarika et al [BPG04] determine if there is a correlation between GRBAS scores and MDVP (noise-related feature). They used thirty-seven patients who are scored by an SLT. A multivariate regression model was used for determining the correlation of these features with GRBAS components. NHR , Voice Turbulance Index (VTI) and Soft Phonation Index (SPI) were reported as the features that have correlation with GRBAS. This can be used for the computerised measurement of GRBAS scores.

2.15 Conclusions

This chapter surveys background knowledge, research literature and software support that is relevant to the work in this thesis. It discusses the nature of speech and the common causes and manifestations of voice impairment. Voice problems are traditionally assessed by the perceptual evaluation of voice features by clinicians. The GRBAS and the CAPE-V approaches standardise these assessments, but they are time-consuming and expensive. Self-assessment has a useful role.

The objective analysis of speech quality has been addressed in the literature and the possibility of producing objective assessments conforming to the GRBAS scale has been realised and investigated to a degree. However the problem has not been solved. There is a wide range of published and commercialised feature measurement software but many authors have reported inconsistencies in the measurements obtained from different software suites. Particular difficulties arise when the aim is to apply them to connected speech as well as sustained vowels. Many different views about which voice features should be measured to detect voice abnormality have been presented in the literature. A wide range of techniques and algorithms for feature measurement have been published or commercialised over the passing years. The nature, inconsistencies, and possible advantages and disadvantages of some of these have been discussed in this Chapter. It is concluded that, based on the knowledge gained from studying this published and commercial software, the means of measuring most of the feature parameters required for predicting GRBAS scores should be developed as part of this thesis. The algorithms developed may be compared and tested against the Praat, MDVP and ADSV software.

We aim to investigate which features are most likely to be indicative of GRBAS scores and which algorithms best serve to measure these features from speech. It may be possible to improve on existing algorithms by tailoring them to the application. There are a number of machine learning techniques available for training classifiers and regression based techniques. Methods of training, validating and testing these are recommended in the machine learning literature. The next chapter is about our database creation and GRBAS scoring by five speech and language therapists (SLTs) and statistical analysis of the GRBAS scores. Chapter 4 will describe new methods for feature measurement.

Chapter 3

Data-base Creation and GRBAS Scoring

3.1 Introduction

This chapter describes the methodology that was used by Gadepalli [C.G13a] for establishing a data-base of voice recordings from 'patients' and 'control' participants. It then explains how the data-base was used for gaining familiarity with traditional subjective GRBAS scoring techniques and for experimental purposes for GRBAS scoring both subjective and objective. Methods of evaluating the 'intra' scorer subjective consistency of individual scorers and 'inter' scorer consistency between different scorers are explored and applied to the data-base. The evaluations have many uses, for example in giving feedback to the SLT scorers. In this thesis, the data-base, the scorings and the scorer evaluations are also used as a means of developing objective GRBAS scoring by computer and assessing the effectiveness of the computerized objective software. The machine learning techniques used to make the objective assessment will be trained using 'training data' consisting of examples extracted from the data-base with GRBAS scores which can be considered 'reliable'. The deviation of the 'reliable' scores from a number of trained subjective scorers will be considered in this Chapter. We refer to these 'reliable' scores as our 'gold standard' GRBAS scores. In this chapter 'N' will consistently denote the number of subject and 'i' will be used to index these subjects. The number of scores will be denoted by small 'n'.

3.2 Data-base creation

Voice data has been collected by Gadepalli [C.G13a] at the Manchester Royal infirmary (MRI) from a random selection of 46 patients and 56 controls [C.G13a]. Ethical approval in the appendix A was obtained by MRI for the data-base and the mode of collecting the voice samples. Only participants that can read English fluently were included in this study. All participants were adults between 18 and 70 years of age, and they were in different stages of their treatment. Information about the participants was stored in secure files. The acoustic signal was captured by a high quality Shure SM48 microphone that was held a constant distance of 20cm from the lips and digitised using the KayPentax 4500 CSL Computerised Speech Laboratory [Kay96]. Each participant was required to sign a consent form after being given an explanation of the nature and purpose of the research. Each recording consists of :

1. Sustained vowel /a/ spoken for about 5 seconds recorded in Mono and Stereo without EGG (Electroglottogram).

2. Sustained vowel /i/ spoken for about 5 seconds recorded in Mono and Stereo without EGG

3. Sustained vowel /a/ spoken for about 5 seconds recorded in Mono and Stereo with EGG

4. Sustained vowel /i/ spoken for about 5 seconds recorded in Mono and Stereo with EGG

5. A set of six standard sentences as specified by CAPE-V (Consensus for auditory perception and evaluation) from a flash card which takes about 12 seconds.

   (a) ‘ The blue spot is on the key again’

   (b) ‘ How hard did he hit him?’

   (c) ‘ We were away a year ago’

   (d) ‘We eat eggs every Easter’

   (e) ‘ My mamma makes lemon jam ‘

   (f) ‘Peter will keep at the peak ‘

6. About 15 seconds of free unscripted speech. Participants speak about something that they like to say for example their daily routines.

3.3 GRBAS scoring

3.3.1 Introduction and motivation

Traditionally, GRBAS scoring is carried out by speech and language therapists (SLTs) interviewing patients, requesting various standard vocal manoeuvres and recording the GRBAS scores in written form. The assessment session may be recorded for future reference. The GRBAS scores may be stored in patient's records in paper or computerised form. This face-to-face form of GRBAS scoring, as widely practised currently, has its advantages resulting from the interaction between patients and SLTs. However, it is time consuming and administratively demanding on staff and patients. An alternative computerised approach, where the patient makes recordings which are to be assessed at a later stage has many advantages over the face-to-face approach though there may be disadvantages due to the loss of face-to-face contact. The computerised recording session can be controlled by a computer program, thus allowing the recording session to be supervised by less highly qualified staff. In some cases, even in clinics, the recording session may not need supervision at all, save to start and stop the session and be on hand in case any failure or misunderstandings occur. In further developments, such recording session could even be carried out by the patient himself/herself at home, after which the recording would be sent electronically to the voice clinic. These developments divide the GRBAS assessment procedure into two parts:

1. Recording session
2. GRBAS assessment session

In building up the database, Gadepalli [C.G13a] has developed and exercised a prototype of the style of recording session that will be recommended in future. The recording sessions carried out by Gadepalli were very carefully standardised and choreographed, following guidelines that are suitable to be adopted in automated recording sessions. The GRBAS assessment session can now be arranged as a private listening session, organised at the convenience of the expert GRBAS assessors. A GRBAS assessor can listen to the recordings of many patients in a single listening session. The possible disadvantages of such listening sessions are that they may prove tiring and they may give different results from what would have obtained in face-to-face sessions. The difference must be borne in mind. But there are clear advantages to be gained in terms of convenience and in other aspects also.

For example, the performance and consistency of scorers may be monitored by the computer software during these listening sessions. The SLTs will enter the scores directly into a computer. This eliminates the pen and paper approach, adds to security and allows assessment to be checked with repeated listening where necessary. It may be questioned whether the result of a live session may be expected to be different from the recorded session in GRBAS scoring. This question has not been addressed in the thesis but would be a useful topic for further research. 'Intra-scorer consistency' can be assessed even when there is just one scorer and this can be useful in providing feedback to the scorer. Where an experiment has the luxury of more than one scorer, 'inter-scorer consistency' can also be assessed and can provide valuable evidence about the likely consistency in GRBAS scoring in general.

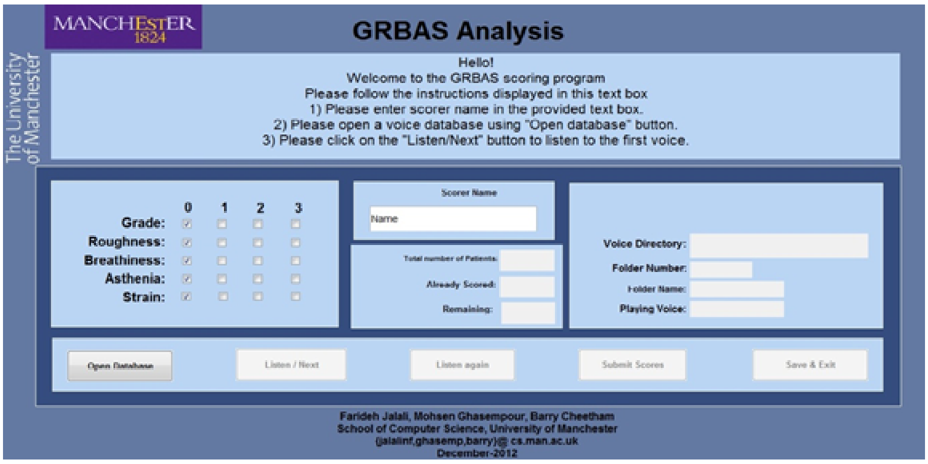
3.3.2 Automated GRBAS scoring

To facilitate the scoring process, we developed a 'GRBAS Presentation and Scoring Package (GPSP)' for collecting GRBAS scores. A graphical user interface (GUI) was created in MATLAB. This interface helps SLTs to store the GRBAS scores in the database, avoids the traditional methods of writing the scores on paper and greatly reduces the risk of losing data. The user interface (GUI) is a graphical display within a window containing control inputs. Each control, and the GUI itself, has call-backs to service the requirements. MATLAB GUIs can be created in two ways: programmatically and using MATLAB's GUIDE software [MAT12]. We used the GUIDE approach which starts with a figure that the programmer populates with components selected using a graphic layout editor. GUIDE creates an associated code file containing callbacks for the GUI and its components. GUIDE saves both the figure (as a FIG-file) and the code file. The application can be launched from either file. The graphical user interface presented by this package is shown in Figure 3.1 [JGA+13]. The GUI is designed to play out in random order and with appropriate repetition, the voice samples from the data-base of recordings. The GUI requires each SLT to follow three instructions to start the GRBAS scoring. The three instructions are as follows.

1. 'Enter the scorer name in the provided text box'.

2. 'Open a voice data-base using the 'Open data-base' button'.

3. Click to the 'Listen/Next' button.



Hello!
Welcome to the GRBAS scoring program
Please follow the instructions displayed in this text box
1) Please enter scorer name in the provided text box.
2) Please open a voice database using "Open database" button.
3) Please click on the "Listen/Next" button to listen to the first voice.

	0	1	2	3
Grade:	☑	□	□	□
Roughness:	☑	□	□	□
Breathiness:	☑	□	□	□
Asthenia:	☑	□	□	□
Strain:	☑	□	□	□

Scorer Name: Name

Total number of Patients:
Already Scored:
Remaining:

Voice Directory:
Folder Number:
Folder Name:
Playing Voice:

Farideh Jalali, Mohsen Ghasempour, Barry Cheetham
School of Computer Science, University of Manchester
{jalalinf,ghasemp,barry}@ cs.man.ac.uk
December-2012

Figure 3.1: This is a Screen shot of the GPSP. SLT scores the GRBAS attributes from 0 to 3. The scorer name in is provided in the text box. Each SLT can observe firstly, the total number recordings that should be scored, secondly, the recordings that have been scored and thirdly the remaining recordings that have not been scored. Five buttons for 'open Datbase', Listen/Next, Listen again, submits scores, and save and exist are provided.

Step 3 causes the GPSP system to select, at random, one of the recordings in the data-base and to play this out to be listened to and scored. The system has a record of which recordings the particular scorer has already completed, though it may choose to repeat one of these to allow the consistency of the scorer to be assessed. The GPSP enables GRBAS scores to be conveniently entered by the SLT and edited if necessary to correct mistakes. The scores are then conveniently recorded in the data-base and may be exported to an excel spread-sheet. By clicking on the 'save and exit' button, the SLT can save his or her GRBAS scores in the data-base along with the name of SLT, and other data. The SLTs are given the option of listening to any samples again by selecting a 'listen again button', and the GUI can be paused at any point, without loss of data. The user may therefore take breaks to prevent tiredness which may affect the scoring.

Each SLT can observe firstly, the total number recordings that should be scored, secondly, the recordings that have been scored and thirdly the remaining recordings that have not been scored. The scoring of the 102 voice examples referred to in this thesis was completed by each SLT in two sessions.

The GUI is user-friendly and is designed be suitable for use in the NHS, Speech-Language-Hearing Association, private hospitals, university hospitals and medical training school for GRBAS scoring.

The 102 recording samples were assessed by five speech and language therapists (SLT) using Sennheiser HD205 head-phones. All the SLTs had been professionally trained for GRBAS scoring and had gained much clinical experience over many years. The 102 voice recording examples were played out in random order with 21 randomly chosen samples repeated as a test for consistency within scorers. In total, 123 voice samples were played out to each SLT. Each voice was scored by each SLT according to the five parameters of the GRBAS scale. For each GRBAS attribute, the possible scores are ‘0’, ‘1’, ‘2’ and ‘3’.

3.4 Reliability testing of GRBAS scoring

Assessing the quality of the GRBAS scoring by health professionals is fundamental not only for clinical care but also for this research. If two SLTs score the same patient under the same circumstances, the two scores can be different for many possible reasons. The SLTs participating in our work were all professionally trained clinicians with experience in the use of GRBAS. In principle, all scorers should give the same scores, but in practice this will not be the case. It may be possible to reduce scorer variability for example by designing the listening conditions and support software (GPSP) well. But it is impossible to eliminate variability. The assessment of scoring agreement is an important issue in this research. It is important to assess how much scores agree when the same SLT repeats the same measurement for a participant (intra-scorer consistency). It is also important to assess how much different SLTs agree when they score the same participants (inter-scorer consistency). If the intra or inter scorers agreement is poor, then the usefulness of the scoring will be limited and may not be considered valid [FLP13].

Bland [Bla, B^{+}00] claimed that the assessment of observer agreement is one of the most difficult areas in the study of clinical measurement. He suggested assessing whether measurements taken on the same subject by different observers may be expected to vary more than measurements taken by the same observer making assessment on different occasions. Pearson correlation, Root Mean Square Error (RMSE), Cohen’s Kappa, Weighted Kappa, Fleiss’ Kappa and Intra-class correlation (ICC) [FLP81] are six different measurements that may be considered for reliability testing [She03, Coh68].

3.4.1 Pearson correlation

The Pearson correlation coefficient is a measure of the strength of the linear relationship between pairs of variables. It is referred to as Pearson's correlation or simply as the correlation coefficient. If the relationship between the two variables is not linear, then the correlation coefficient does not adequately represent the strength and nature of the relationship between the variables.

The symbol for Pearson's correlation is 'ρ' (rho) when it is measured in the population and 'r' when it is measured in a sample or subset of the population. We will be dealing almost exclusively with samples, and will use r to represent Pearson's correlation. Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship (i.e. a linear relationship with negative slope) between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship (i.e. a linear relationship with positive slope) between variables [She03]. Table 3.1 depicts an interpretation of the significance of Pearson correlation [Deb].

Value of the Correlation Coefficient (r)	Significance
1	Perfect positive linear correlation
0.8 to 1	Very Strong positive linear correlation
0.5 to 0.8	Strong positive linear correlation
0.3 to 0.5	Moderate positive linear correlation
0.1 to 0.3	Weak positive linear correlation
0	No linear correlation
-0.1 to -0.3	Weak negative linear correlation
-0.3 to -0.5	Moderate negative linear correlation
-0.5 to -0.8	Strong negative linear correlation
-0.8 to -1	Very Strong negative linear correlation
-1	Perfect negative linear correlation

Table 3.1: Significance of Pearson Correlation Coeffs according to [Deb]

Equation (3.1) defines the Pearson correlation [She03] between the two dimensions of a sample $\{(x_\text{i}, y_\text{i})\}$ containing n pairs of random variables (x_i, y_i); $\bar{x}$ and $\bar{y}$ are the sample means of $\{\text{x}_\text{i}\}$ and $\{\text{y}_\text{i}\}$ respectively. Then, a formula for r is:

$$r = \frac{\sum_{i=1}^{n} (x_\text{i} - \bar{x})(y_\text{i} - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_\text{i} - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_\text{i} - \bar{y})^2}} \tag{3.1}$$

Table 3.2 illustrates a fictitious example of possible scores that two scorers may have given for 10 participants. In this example, Scorer 2 consistently scores each participant higher than Scorer 1. In fact, the score given by Scorer 2 is always one division greater than that given by Scorer 1. This means maximum positive (perfect) correlation. The Pearson correlation between scores given by Scorer 1 and the scores given by Scorer 2 is '1'. However, it is clear that there is not a perfect agreement between them. This is because, in Pearson correlation, each variable is centered and scaled by its own mean and standard deviation. If the means for all scorers are the same, Pearson correlation can be a good indicator of absolute agreement. If the means are not the same, it can be misleading if incorrectly interpreted. This illustrates that Pearson Correlation is not necessarily a good way of comparing the consistency of scores.

Participant	Scorer1	Scorer2
1	0	1
2	0	1
3	0	1
4	1	2
5	1	2
6	2	3
7	1	2
8	1	2
9	0	1
10	2	3

Table 3.2: Illustration of possible scores for Scorer 1 and Scorer 2

3.4.2 Root Mean Square Error

The Root Mean Square Error (RMSE) can measure the difference between two sets of scores. In our application, RMSE is used for measuring the difference between the scores of two scorers. Equation (3.2) defines the RMSE [CD14] between the two dimensions of a sample $\{(x_{\mathrm{i}}, y_{\mathrm{i}})\}$ containing n pairs of variables $(x_{\mathrm{i}}, y_{\mathrm{i}})$.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{\mathrm{i}} - y_{\mathrm{i}})^2} \tag{3.2}$$

For table 3.2, the value of RMSE is non-zero despite the fact that the Pearson correlation is 1. If the means of two sets of measurements are not equal, Intraclass Correlation

(ICC), Cohen's kappa, Weighted Kappa and Fleiss' Kappa may be preferred to Pearson correlation for comparing the consistency of scorers.

3.4.3 Intra-Class Correlation (ICC)

ICC assesses the reliability of scorings by comparing the variability of different scorings of the same subject to the total variation across all scorings and all subjects [SF79]. Despite its name, ICC can be used for assessing inter-rater and also intra-rater reliability. A key difference between Pearson Correlation and ICC is that ICC uses a pooled mean and standard deviation, whereas in the Pearson Correlation, each variable is centred and scaled by its own mean and standard deviation. The Intra-class Correlation Coefficient (ICC) is suitable for ordinal continuous or discrete data. Comparing with equation (3.1) the ICC coefficient r for two sets of data $\{x_\mathrm{i}\}_{1,\mathrm{N}}$ and $\{y_\mathrm{i}\}_{1,\mathrm{N}}$ as may have been generated by scorers is given by:

$$r = \frac{\sum_{i=1}^{N} (x_\mathrm{i} - m)(y_\mathrm{i} - m)}{0.5\left(\sum_{i=1}^{N} (x_\mathrm{i} - m)^2 + \sum_{i=1}^{N} (y_\mathrm{i} - m)^2\right)} \tag{3.3}$$

where m is the arithmetic average of the sample-means of $\{x_\mathrm{i}\}$ and $\{y_\mathrm{i}\}$. Therefore ICC measures variation in both $\mathrm{x_i}$ and $\mathrm{y_i}$ about the same mean m, whereas Pearson's correlation compares variation in $\{x_\mathrm{i}\}$ about $\bar{x}$ and $\{y_\mathrm{i}\}$ about $\bar{y}$. Another difference lies in the denominator which is the geometric mean of the two variances for Pearson, and is the arithmetic mean for ICC. The significance of the denominator difference is subtle and not likely to be important for reliability measurement. Applying this formula to the data shown in Table (3.2) produces a much lower value of correlation r than was obtained using Pearson Correlation. The value of ICC obtained for Table (3.2) is 0.38. The ICC formula above may be generalised to three or more scorers to obtain a parameter that indicates the degree of consistency among a group of scorers. To do this, m becomes the arithmetic average of the sample-means obtained for the multiple scorers. To illustrate the generalisation by quoting the formula for three scorers with data $\{x_\mathrm{i}\}$ and $\{y_\mathrm{i}\}$ and $\{z_\mathrm{i}\}$ we obtain:

$$r = \frac{\sum_{i=1}^{N}(x_\mathrm{i}-m)(y_\mathrm{i}-m)+\sum_{i=1}^{N}(x_\mathrm{i}-m)(z_\mathrm{i}-m)+\sum_{i=1}^{N}(y_\mathrm{i}-m)(z_\mathrm{i}-m)}{\left(\sum_{i=1}^{N}(x_\mathrm{i}-m)^2+\sum_{i=1}^{N}(y_\mathrm{i}-m)^2+\sum_{i=1}^{N}(z_\mathrm{i}-m)^2\right)} \tag{3.4}$$

Note that the numerator is built up from pair-wise comparisons of scoring vectors. Three scorers give three pairs, four scorers give 6 pairs, five scorers give ten pairs and so on. Both numerator and denominator must be made arithmetic means of the individual terms. Like Pearson correlation, ICC can only be applied to ordered numerical data which may be continuous or discrete. Therefore ICC can be applied to GRBAS scoring for quantifying both intra-scorer and inter-scorer consistency. For Table 3.2, the value of ICC is ‘0.38’ despite the fact that the Pearson correlation is 1. Both Pearson correlation and ICC are applicable only to ordinal numeric data and automatically weight the contribution of each data item according to its difference from the mean. Table 3.3 shows the significance of ICC values.

ICC	**Strength of Agreement**
$>$ 0.75	Excellent
0.75-0.40	Agreement between Fair and Good
<0.4	Poor

Table 3.3: Significance of ICC values

Different versions of ICC have been proposed, and there has been much discussion about which version is appropriate for a given application. They may produce significantly different results for the same data [MB94, MW96]. The version quoted above is the original pair-wise version which is restricted to applications where the identities and characteristics of all scorers remain constant and all scorers score all subjects.

3.5 Kappa measurements

Kappa statistics was introduced by Cohen to provide a coefficient of agreement between the categorical scores produced by two scorers [C$^+$60]. Categorical or ‘nominal’ scores are scores which are non-numeric (e.g. good, bad, etc.) or where any numeric notation is used just as labels. The significance of categorical scores is only whether they are the same or different [C$^+$60]. There is no concept of ‘greater than’ and ‘magnitude of difference’ with categorical scoring. Where scores are actually numerical, with grades of difference as with GRBAS, they can be considered as categorical by simply considering the numerical scores as labels. But there is loss of possibly useful information in doing this.

The original definition of Kappa presented by Cohen in 1960 [C$^+$60] is suitable for comparing just two raters where there are two or more nominal response categories. Cohen later introduced the ‘weighted Kappa’ [Coh68] which is applicable to ordered numeric data rather than categorical data. Weighted Kappa allows any disagreement between two scorers to be weighted according to the degree of the disagreement. There are now ordinal categories which may or may not be numerical. For example, in medical terminology, a disagreement ‘suspicious’ and ‘normal’ may be considered to represent a less strong disagreement than a disagreement between ‘pathological’ and ‘normal’. The equivalence of Cohen’s weighted Kappa to ICC (discussed in the previous section) under some circumstances was shown by Fleiss [Fle71]. Fleiss also proposed an extension to Cohen’s original Kappa allowing the assessment of agreement between several scorers that score all or some of the subjects and classify them into two or more categories [Fle71]. So far, the Fleiss Kappa has been applicable only to categorical data, but in this thesis we propose a ‘weighted Fleiss Kappa’ which is applicable to ordinal data.

3.5.1 Cohen’s Kappa

Given two sets of scores for the same set of N patients, a Score-Distribution matrix as in Table 3.4, may be constructed to show the number of patients which scorer A scores as category i and scorer B scores as category j. There can of course be more than four categories. From this, a simple measure of the consistency of scoring may be readily deduced as the proportion of subjects for which the two scorers agree. It is termed ‘$\mathrm{p_o}$’ and is the sum of the diagonal terms of the matrix divided by N which is 27 in this example. Therefore $p_\mathrm{o} = 16/27$.

However there is a difficulty with this simple measure since there will always be a probability of agreement by chance even if the scoring has been done arbitrarily without reference to the subjects. Cohen’s Kappa estimates the probability of agreement by chance given the distribution of scores produced by Scorers A and B. This is easily done from Table 3.4 by estimating the probability of scorers A and B both producing category 0 given only their observed distributions of scores. Since Scorer

		Scorer B			
		Cat 0	Cat 1	Cat 2	Cat 3
Scorer A	Cat 0	5	1	1	0
	Cat 1	3	6	2	1
	Cat 2	1	1	2	0
	Cat 3	0	0	1	3

Table 3.4: Score-distribution table

A produces category 0 seven times out of 27 and Scorer B produces category 0 nine times out of 27, the probability of both producing category 0 at the same time, by chance, is $(7/27) \times (9/27)$. This is repeated for categories 1, 2 and 3, and then all four joint probabilities are summed to obtain an estimate of the probability of agreement by chance.

The Cohen Kappa coefficient is defined by equation (3.5) where $\mathrm{p_o}$ is the proportion (between 0 and 1) of subjects for which the two SLTs agree (exactly) on the scoring, and $\mathrm{p_e}$ is the probability of agreement ‘by chance’.

$$\kappa = \frac{p_{\mathrm{o}} - p_{\mathrm{e}}}{1 - p_{\mathrm{e}}} \tag{3.5}$$

Kappa is widely used for comparing raters or scorers, and reflects any consistent bias in the average scores for each scorer [VG^{+}05]. Table 3.5 shows how K values in different ranges may be interpreted, according to [LK77]. Cohen’s Kappa takes the scores to be ‘categories’ rather than ordinal numerical data. Only ‘agreement’ and ‘disagreement’ is taken into account. The degree of disagreement in ordinal numerical scoring is disregarded. Applying this to GRBAS scoring with categories 0, 1, 2 and 3, if Scorer A scores 1, the effect of scorer B scoring 0, 2 or 3 will be essentially the same. The fact that 3 is a worse discrepancy than 2 is disregarded. Cohen’s Kappa is applicable when each rater completes the task of scoring all the subjects. The scoring characteristics of each rater determine the probability of agreement by chance. Some measurements of reliability allow the identity of raters to change during the scoring process and do not need each patient to be scored by each scorer. Cohen’s Kappa requires all patients to be rated by each scorer and assumes that the characteristics of each scorer do not change during the scoring process.

K	Agreement
$\leqslant 0$	Poor
0.01-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1	Almost Perfect

Table 3.5: Significance of Kappa values

3.5.2 Weighted Kappa

Weighted Kappa is often more appropriate when there are more than two possible 'ordinal' numerical scores with a sense of distance between the scores [Coh68]. With possible scores 0, 1, 2, 3, Kappa only considers agreement or disagreement between scorers, whereas weighted Kappa takes into account the degree of disagreement. In this application, discrepancy between scores 0 and 2, for example, is more serious than the difference between 0 and 1 or between 1 and 2, and weighted Kappa takes this into account. With linearly weighted Kappa, the disagreement between 0 and 2 may be weighted twice that between 0 and 1, 1 and 2, or 2 and 3. The discrepancy between 0 and 3 may be weighted three times that between 0 and 1. When there are c = 4 scoring categories, this linear weighting of discrepancy is conveniently expressed by a *c by c* 'weighting matrix' W_{L} as follows:

$$W_{\mathrm{L}} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{bmatrix} \tag{3.6}$$

It is useful to define a 'proportion distribution' (PD) matrix $\mathrm{P_D}$ as follows:

$$P_{\mathrm{D}} = (1/N)S_{\mathrm{D}} \tag{3.7}$$

where $\mathrm{S_D}$ is the *k by k* 'scoring distribution' matrix with entries as illustrated by Table (3.4) and scalar constant N is the total number of subjects. Element (i, j) of $\mathrm{P_D}$ is the proportion (between 0 and 1) of N subjects which were scored i by scorer A and scored j by Scorer B. For GRBAS scoring, $0 \leqslant i \leqslant 3$ and $0 \leqslant j \leqslant 3$. The entries of P_{D} are termed 'proportions' rather than 'probabilities' because they reflect actual results from the scoring rather than expectations of likely scorings.

We also define a $k \times k$ 'expectation matrix' E as follows:

$$E = \begin{bmatrix} p_{\mathrm{A0}}p_{\mathrm{B0}} & p_{\mathrm{A0}}p_{\mathrm{B1}} & p_{\mathrm{A0}}p_{\mathrm{B2}} & p_{\mathrm{A0}}p_{\mathrm{B3}} \\ p_{\mathrm{A1}}p_{\mathrm{B0}} & p_{\mathrm{A1}}p_{\mathrm{B1}} & p_{\mathrm{A1}}p_{\mathrm{B2}} & p_{\mathrm{A1}}p_{\mathrm{B3}} \\ p_{\mathrm{A2}}p_{\mathrm{B0}} & p_{\mathrm{A2}}p_{\mathrm{B1}} & p_{\mathrm{A2}}p_{\mathrm{B2}} & p_{\mathrm{A2}}p_{\mathrm{B3}} \\ p_{\mathrm{A3}}p_{\mathrm{B0}} & p_{\mathrm{A3}}p_{\mathrm{B1}} & p_{\mathrm{A3}}p_{\mathrm{B2}} & p_{\mathrm{A3}}p_{\mathrm{B3}} \end{bmatrix} \tag{3.8}$$

where p_{Ai} is the probability of scorer A scoring 'i' by chance as estimated from the distributions of scores from scorer A. Similarly p_{Bj} is the estimated probability of scorer B scoring 'j' by chance as estimated from B's distribution of scores.

$$P_{\mathrm{Ai}} = \sum_{j=0}^{c-1} P_{\mathrm{D}}(i,j) \qquad P_{\mathrm{Bj}} = \sum_{i=0}^{c-1} P_{\mathrm{D}}(i,j) \tag{3.9}$$

Summing the off-diagonal elements of E gives the probability of disagreement by chance. This may be considered as a cost arising from this probability of disagreement by chance. We call this scalar $\mathrm{D_e}$. Similarly, summing the off-diagonal elements of $\mathrm{P_D}$ gives the cost of disagreement in the actual scorings. Call this scalar $\mathrm{D_o}$. It follows that the probability of agreement by chance $\mathrm{p_e}$ is given by:

$$p_{\mathrm{e}} = 1 - D_{\mathrm{e}} \tag{3.10}$$

Similarly, the proportion of agreement in the actual scorings is:

$$p_{\mathrm{o}} = 1 - D_{\mathrm{o}} \tag{3.11}$$

Finally, the weighted Cohen Kappa can be expressed as:

$$\kappa = \frac{p_{\mathrm{o}} - p_{\mathrm{e}}}{1 - p_{\mathrm{e}}} = \frac{D_{\mathrm{e}} - D_{\mathrm{o}}}{D_{\mathrm{e}}} = 1 - \frac{D_{\mathrm{o}}}{D_{\mathrm{e}}} \tag{3.12}$$

This is essentially the Cohen's Kappa but now expressed in terms of disagreement rather than agreement. Expressing the Cohen Kappa in this way is very convenient because it will allow the actual disagreements and probabilities of disagreements by chance to be weighted according to numerical differences.

Performing element by element multiplication between the '$\mathrm{P_D}$ matrix' and the 'Expectation matrix' E, by element of $\mathrm{W_L}$, changes the cost of disagreement depending on how different the scores are. If $\mathrm{W_L}$ were replaced by $\mathrm{W_U}$ or $\mathrm{W_S}$ as defined below different weighting would be obtained. If the weighting is by $\mathrm{W_U}$ (U for unweighted), the weighting would be exactly as in the original Cohen Kappa with any discrepancy equally weighted. If the weighting is by W_S this is non-linear 'squared' weighting where discrepancies of 2 and 3 in scores would cost, respectively, four and nine times a discrepancy of one.

$$W_U = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \qquad W_S = \begin{bmatrix} 0 & 1 & 4 & 9 \\ 1 & 0 & 1 & 4 \\ 4 & 1 & 0 & 1 \\ 9 & 4 & 1 & 0 \end{bmatrix} \tag{3.13}$$

From Equation (3.12), it can be seen that this weighting approach leads to Equation (3.14) for linearly weighted Kappa, where P_D(i,j) denote the elements of matrix P_D. Therefore, P_D(i,j) is the proportion of subjects that are scored i by scorer 1 and j by scorer 2. Element E(i,j) of matrix E is the 'by chance' probability of scorer 1 scoring i while scorer 2 scores j. Matrix E was estimated from the observed distribution of scores by each scorer, but with no correlation between scorers. The number of scoring categories is c.

$$\kappa w = 1 - \frac{\sum_{i=0}^{c-1} \sum_{j=0}^{c-1} W_L(i,j) P_D(i,j)}{\sum_{i=0}^{c-1} \sum_{j=0}^{c-1} W_L(i,j) E(i,j)} \tag{3.14}$$

Table 3.5 is commonly assumed [LK77] as the significance of Weighted Kappa as well as the un-weighted form. Replacing W_L in Equation (3.14) by W_U gives ordinary Cohen's Kappa and other weighting matrices, such as W_S in Equation 3.13 could be used instead.

3.5.3 Fleiss' Kappa

Cohen's Kappa is applicable only when there are just two scorers and where the same two scorers score every subject. Fleiss's Kappa [Fle71, FLP81] measures agreement among any number of scorers. It also caters for case where the scorers for each subject may be different, although this case is not of interest in this thesis. Like Cohen's original un-weighted Kappa, Fleiss Kappa considers only agreement or disagreement, and treats scores as categories rather than ordinals. To explain Fleiss Kappa, it is convenient to concentrate initially on just one of the GRBAS properties; we chose 'Grade'. If N is the number of subjects, n is the number of scorers and the possible 'Grade' categories are indexed 0, 1, 2 and 3, Fleiss [FLP81] would define n_{ij} as the number of scorers who give subject i a 'Grade' score of j. Clearly, for each subject i,

$$\sum_{j=0}^{3} n_{\mathrm{ij}} = n \tag{3.15}$$

because each subject is scored by n scorers. To illustrate this point, if patient 1 is scored 0, 0, 0, 1, 2 by 5 scorers, then $n_{10} = 3$, $n_{11} = 1$, $n_{12} = 1$ and $n_{13} = 0$. Now define pj to be the proportion of all assignments, across all subjects and all scorers, to GRBAS score j. Then:

$$p_{\mathrm{j}} = \frac{1}{nN} \sum_{i=1}^{N} n_{\mathrm{ij}} \qquad \text{and} \qquad \sum_{j=0}^{3} p_{\mathrm{j}} = 1 \tag{3.16}$$

Fleiss Kappa quantifies the extent of agreement among the n scorers for subject i as the proportion of pairs of scorers that agree for that subject. This proportion is:

$$P_{\mathrm{i}} = \frac{1}{n(n-1)/2} \sum_{j=0}^{3} n_{\mathrm{ij}}(n_{\mathrm{ij}} - 1)/2 \tag{3.17}$$

since the number of non-ordered pairs of objects out of a total of n is n(n-1)/2. The overall agreement across all subjects may now be measured by the average of the $\mathrm{P_i}$ values, i.e.

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_{\mathrm{i}} \tag{3.18}$$

Fleiss explains that the significance of $\bar{P}$ is that if subject 'i' is scored by two scorers chosen at random from the n available scorers, the probability that the scores will agree is $\bar{P}$. This may be taken as an estimate of inter-scorer consistency. However, some degree of agreement will always occur by chance and Fleiss Kappa, like Cohen's Kappa, tries to factor this chance agreement out of the estimate. It is argued that if all scorers made their assignments purely at random with respect to the subjects, but with the same distribution of scores as observed in the actual scoring exercise, the probability of agreement between pairs may be estimated as:

$$P_{\mathrm{e}} = \sum_{j=0}^{3} {p_{\mathrm{j}}}^2 \tag{3.19}$$

This is the probability of getting score j twice, by chance, summed over the 4 possible GRBAS scores. The maximum probability of agreement among pairs which is not by chance is therefore $1 - P_{\mathrm{e}}$. It is less than 1 because of the 'by chance' element. The degree of agreement actually achieved in the scoring procedure which is not due to pure chance is estimated as $\bar{P} - P_{\mathrm{e}}$. Therefore the normalised measure of agreement among the n scorers which constitutes the Fleiss Kappa is as follows:

$$\kappa = \frac{\bar{P} - Pe}{1 - Pe} \tag{3.20}$$

The Fleiss' Kappa coefficient was used to measure the level of agreement and inter-scorer consistency between five SLT scorers [FLP81] for Grade and subsequently for the other GRBAS parameters. The results will be presented later.

3.5.4 Fleiss Kappa demystified

Fleiss Kappa is not easy to understand. It may be shown to be a generalisation of the Cohen's Kappa, in that calculation of Fleiss Kappa for two scorers does give the Cohen's Kappa. It is similar to ICC in considering all possible pairs of scorers, but is currently applicable only to data that is considered categorical (rather than ordinal). There is no published form of weighted Fleiss Kappa.

It may be considered that a reasonable alternative to the Fleiss Kappa for n scorers would be the average of $n(n-1)/2$ unweighted Cohen Kappa scores covering every possible pair of scorers. When there are n=5 scorers for 'Grade', this would be the average of ten unweighted Cohen Kappas. The average value of $\mathrm{P_o}$ obtained from the ten unweighted Cohen Kappa calculations is always identical to $\bar{P}$ in Equation 3.18 regardless of how the scores are distributed. Also, where all four 'Grade' scores are equally probable (with p1 = p2 = p3 = p0 = 0.25 for Grade), the Fleiss Kappa is identical to the average of the ten Cohen Kappas for Grade. Similarly for R, A, B and S. Where the distribution of scores is not evenly spread, the values of $\mathrm{P_e}$ will vary for Cohen Kappa from pair to pair, and ultimately there will be small difference between the average Cohen Kappa and the Fleiss Kappa. In practice the differences are rarely large, and it may be argued that calculating the probability of agreement/disagreement by chance is better done on a pair by pair basis than on the basis of all scorers at once. It is possible that the average Cohen Kappa is equally acceptable or even preferable to the unweighted Fleiss version.

Even more significantly, this argument can lead us to a definition of a form of weighted Fleiss Kappa, simply by averaging weighted Cohen Kappa values between all possible pairs of scorers. This was not the approach taken in this thesis to develop a weighted Fleiss Kappa, though it was later realised that the approach actually taken (described below) is identical to the average of weighted Cohen Kappas when distributions of scorings are even (p0 = p1 = p2 = p3 = 0.25). In practice, differences remain small for uneven distributions, and the same arguments for averaged weighted Cohen Kappas against the new weighted form of Fleiss Kappa (to be presented next) can be made as for the non-weighted case.

It is useful to consider the possible values $\mathrm{P_i}$ for traditional Fleiss for each subject i when there are five scorers and four Grade categories: 0, 1, 2 and 3. The same arguments apply to the other GRBAS parameters and we single out Grade just to simplify the explanation.

Agreement	**Examples**	**Pi**
All 5	0,0,0,0,0	10/10=1
4 out of 5	0,0,0,0,1 or 0,3,3,3,3	6/10=0.6
3 and 2 out of 5	0,0,0,1,1 or 0,0,0,3,3	(3+1)/10=0.4
3 out of 5 only	0,0,0 ,1,2 or 0,0,0,2,3	3/10=0.3
2 out of 5 only (twice)	0,0,1,1,2 or 0,0,2,3,3	(1+1)/10
2 out of 5 only	0,0,1,2,3 or 0,1,2,3,3	1/10
none	impossible	n/a

Table 3.6: Examples of Fleiss Kappa.

Fleiss Kappa works by taking $(1-P_\mathrm{i})$ as a cost for each subject and averaging this cost over all N subjects. This clearly has validity, but also a degree of arbitrariness in the cost that becomes associated with each type of disagreement. For example the cost of the disagreement when the five scorers score 0, 0, 0, 0, 1 is 1 - 0.6 = 0.4 which is exactly the same as for 0, 3, 3, 3, 3. Ideally, the measure of consistency should be higher for the first of these two examples than for the second. In other words, the cost of inconsistency for the first of these two examples should be lower than that for the second. It also clear from the examples in the Table 3.6 that scoring we may consider to have quite different significance, such as 0, 0, 0, 1, 1 and 0, 0, 0, 3, 3 are given the same measure of consistency which is derived directly from the number of matching pairs of scores.

This table makes it clear that different cost weightings are perfectly possible and do not have to be derived purely from the proportions of scorings. In the next section we propose a new form of Fleiss Kappa which may be weighted according to the application. The designer can define his/her own weighting either ad-hoc or according to some algorithm like linear or squared weighting. This has obvious advantages, but also the disadvantage of being new and non-standard. The Fleiss Kappa, ideal or not, is well known and well used throughout the literature, whereas a new weighted Fleiss Kappa would have to be justified and argued for. The new approach may be applied to other applications with any number of scorers and possible scores. In this Chapter, we restrict it to the current application of GRBAS scoring, concentrating initially on ‘Grade’.

3.5.5 Weighted Fleiss Kapppa and Farideh’s Kappa

A new weighted form of Fleiss Kappa may be devised simply by replacing the $\mathrm{P_i}$ column in Table (3.7) by some other parameter $\mathrm{Q_i}$ which is no longer the proportion of matching pairs of scores. In fact $\mathrm{Q_i}$ may be chosen by the designer arbitrarily in the range 0 to 1 to reflect the degree of agreement considered to exist among the five scorers. Define the Fleiss ‘linear weighting vector’ as:

$$W_{\mathrm{L}} = \begin{bmatrix} 0 & 1 & 2 & 3 \end{bmatrix} \tag{3.21}$$

and the Fleiss ‘unweighting vector’ as:

$$W_{\mathrm{U}} = \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix} \tag{3.22}$$

Let $\mathrm{D_{ij}}$ denote the number of scorers pair that differ by j (for j=0,1,.., 3) for subject i. For example, if the scores for subject 1 are 0 0 0 0 0, then $\mathrm{D_{10}}$ =10, $\mathrm{D_{11}}$=0 $\mathrm{D_{12}}$=0 $\mathrm{D_{13}}$=0. If the scores for subject 2 are 0 0 0 1 1 then $\mathrm{D_{20}}$ =4, $\mathrm{D_{21}}$ =6, $\mathrm{D_{22}}$=0 & $\mathrm{D_{23}}$=0. If the scores for subject 3 are 0 1 2 3 3 then $\mathrm{D_{30}}$ = 1, $\mathrm{D_{31}}$= 4, $\mathrm{D_{32}}$=3, $\mathrm{D_{33}}$=2. An algorithm for calculating the D matrix is easily derived. Let $\mathrm{N_p}$ equal the number of pairs which for 5 scorers is $5 \times 4/2 = 10$. If, for each subject i, we define the ‘degree of (actual) agreement’ $\mathrm{Q_i}$ as:

$$Qi = 1 - (1/N_{\mathrm{p}}) \sum_{j=0}^{3} W(j) D_{\mathrm{ij}} \tag{3.23}$$

with W = $\mathrm{W_U}$, then $\mathrm{Q_i}$= $\mathrm{P_i}$ as calculated for traditional Fleiss Kappa by equation (3.17). Similarly, defining $\mathrm{p_j}$ as in Equation (3.16) for traditional Fleiss Kappa, we can obtain the ‘degree of (probable) pair agreement’ by chance as:

$$Qe = 1 - \sum_{j=0}^{3} \sum_{c=0}^{3} W(|j-c|) p_{\mathrm{j}} p_{\mathrm{c}} \tag{3.24}$$

With W=W_U, this Equation gives Q_e which is exactly equal to P_e as given for traditional Fleiss Kappa by Equation (3.19). Therefore we can calculate traditional Fleiss Kappa in this revised way, with the 'unweighted' weighting matrix given by Equation (3.22). Replacing W in Equations (3.23) and (3.24) by W_L as given by Equation (3.21) gives a new unpublished linearly weighted form of Fleiss Kappa which reflects the differences between different pairs of scores. Its Equation is:

$$\kappa = \frac{\bar{Q} - Qe}{1 - Qe} \tag{3.25}$$

where $\bar{Q}$ is the average of Q_i over all subjects and Q_e is defined by equation (3.24) with weighting vector W=W_L. It follows that the new Fleiss Kappa is given by

$$\kappa = 1 - \left(\frac{(1/(N \times N_p)) \sum_{i=1}^{N} \sum_{j=0}^{3} W(j) D_{ij}}{\sum_{j=0}^{3} \sum_{c=0}^{3} W(|j-c|) p_j p_c} \right) \tag{3.26}$$

where N is the number of subjects and p_j denotes the proportion of all assignments, across all subjects and all scorers, to Grade score j. The original and new linear weightings given by the new Fleiss Kappa, adapted to the GRBAS application, are illustrated by the values of P_i and Q_i in the table given below.

Agreement	**Examples**	**Pi**	**Qi**
All 5	0,0,0,0,0	1	1
4 out of 5	0,0,0,0,1	0.6	0.6
	0,3,3,3,3	0.6	-0.2
3 and 2 out of 5	0,0,0,1,1	0.4	0.4
	0,0,0,3,3	0.4	-0.8
3 out of 5 only	0,0,0,1,2	0.3	0
	0,0,0,2,3	0.3	-0.6
2 out of 5 (twice)	0,0,1,1,2	0.2	0
	0,0,1,3,3	0.2	-0.8
2 out of 5 only	0,0,1,2,3	0.1	-0.6
	0,1,2,3,3	0.1	-0.6
none	impossible	n/a	

Table 3.7: The example of original and new linear weightings given by the new Fleiss Kappa.

The negative values of Q_i occur in the original weighted Cohen's Kappa and also in the new weighted Fleiss Kappa. They highlight the fact that the weighted cost is no longer a probability, and it need not be.

The linear weighting and the table it produces may be considered somewhat arbitrary. The values of $\mathrm{Q_i}$ are intended to reflect both the number of scorers in agreement and the severity of any disagreement. Clearly costings different from those implicit in traditional Fleiss Kappa have been used. But there are many other costing formulae that could have been applied. In practice, especially with only 5 scorers, a table of costings may be derived for all possible combinations of scores. Such a table would only have 1024 rows if expressed in its most inefficient (and non-scalable) form. In such a form yet more advantages emerge, such as the possibility of considering agreement for higher scores (2 and 3 say) as more valuable than agreement for score 0. Hence the disagreement in 2, 3, 3, 3, 3 could be defined as costing less than that in 0, 0, 0, 0, 1. A table defined in this way has been termed the 'Farideh Kappa' which then defines $\bar{Q}$ as the average of $\mathrm{Q_i}$ over all subjects instead of the average of $\mathrm{P_i}$. We may present more on this later.

3.5.6 Intra-Scorer Consistency

'Intra-Scorer Consistency' is the agreement between the scores given by a single scorer when he or she has been required to score some of the samples twice, ideally without realising. About 20% of the recordings were played out twice to each SLT. These were chosen at random independently for each scorer, which meant that each scorer repeated a different set of recordings. The GPSP GUI was programmed to play out the required small number of repeated examples interspersed with non-repeated samples to try to divert attention from the fact that some samples are being repeated. Repeated samples were required to be re-graded in all five GRBAS categories and without reference to the previous gradings. Therefore, 'intra-scorer consistency' was investigated for each scorer using GRBAS re-scoring for about 20 participants.

Pearson correlation, Cohen's Kappa, Weighted Cohen Kappa and RMSE were considered for testing 'intra-scorer consistency' for each of the five SLT scorers. Firstly, the means of the repeated scores were computed along with the means of the original versions of these scores. Table 3.8 presents the mean of each GRBAS component for the first 20 of the randomly chosen repeated samples before and after they were re-scored.

For the GRBAS attributes that do not have equal means for the first and second times of scoring, the Pearson correlation is not a good measurement for testing the consistency. Kappa and Weighted Kappa are preferable to the Pearson Correlation in

	G		R		B		A		S	
SLT	Orig	Rep	Orig	Rep	Orig	Rep	Orig	Rep	Orig	Rep
1	**0.85**	**0.85**	0.75	0.70	**0.50**	**0.50**	**0.55**	**0.55**	0.45	0.40
2	1.05	1.2	0.90	1.1	0.60	0.70	0.20	0.30	0.45	0.35
3	0.95	1.05	0.40	0.50	0.30	0.75	**0.60**	**0.60**	0.75	0.70
4	1	1.20	0.65	0.75	0.55	0.80	0.65	0.45	0.25	0.20
5	1.30	1.26	0.95	0.60	0.43	0.65	0.52	0.43	0.73	0.34

Table 3.8: Mean of GRBAS scores for 20 random examples scored twice

these cases. To investigate the Kappa and Weighted Kappa measurements, the number of differences between the original and repeated scores was computed for each scorer. It was found that, among all SLT scorers, SLT1 had the lowest number of different scores in Grade and Breathiness. SLT2 and SLT4 had a better confidence than others in Asthenia and Strain scoring. A maximum scoring difference of one was observed for the majority of the SLTs for each GRBAS component. An occasional difference of two occurred, but there were no differences of 3.

Tables 3.9, 3.10, 3.11, 3.12 and 3.13 compare the number of differences in scoring, Pearson Correlation, ICC, Cohen's Kappa, Weighted Kappa and RMSE values as indicators of intra-scorer agreement or consistency. The Pearson correlation values can be considered reliable as indicators of consistency in Grade, Breathiness and Asthenia for SLT scorer 1 since the means of the original and repeated scoring were found to be very close. The Pearson correlation values for Grade, Breathiness and Asthenia are 0.92, 0.88 and 0.77 respectively. SLT3 has the same mean in Asthenia scoring and the Pearson correlation between the original and repeated Asthenia scores is 0.85. For the other SLT scorers and GRBAS components, the Pearson Correlation cannot be a reliable measurement of consistency due to the mean being different in the first and second time of the scoring.

SLT	Different scores	Pearson Corr	ICC	Kappa	Weighted Kappa	RMSE
1	2	0.92	0.92	0.84	0.88	0.31
2	5	0.90	0.89	0.65	0.78	0.50
3	6	0.84	0.84	0.56	0.71	0.54
4	4	0.93	0.90	0.71	0.82	0.44
5	10	0.75	0.73	0.41	0.62	0.82

Table 3.9: Intra-Scorer Consistency in Grade

According to Tables 3.9, 3.10 and 3.11, SLT1 has 'almost perfect agreement' for 'Grade', 'Roughness' and 'Breathiness' according to Weighted Kappa and almost the same agreement for 'Grade' and 'Breathiness' as measured by Cohen's Kappa. All SLT scorers have obtained 'substantial agreement' in 'Asthenia' scoring by Weighted

SLT	Different scores	Pearson Corr	ICC	Kappa	Weighted Kappa	RMSE
1	3	0.89	0.88	0.76	0.82	0.38
2	6	0.88	0.86	0.57	0.73	0.54
3	6	0.51	0.49	0.41	0.44	0.54
4	4	0.86	0.83	0.67	0.75	0.42
5	9	0.72	0.64	0.34	0.50	0.77

Table 3.10: Intra-Scorer Consistency in Roughness

SLT	Different scores	Pearson Corr	ICC	Kappa	Weighted Kappa	RMSE
1	2	0.88	0.88	0.81	0.84	0.31
2	4	0.88	0.87	0.65	0.77	0.44
3	9	0.86	0.62	0.19	0.42	0.67
4	6	0.77	0.72	0.50	0.61	0.67
5	4	0.87	0.83	0.58	0.74	0.59

Table 3.11: Intra-Scorer Consistency in Breathiness

SLT	Different scores	Pearson Corr	ICC	Kappa	Weighted Kappa	RMSE
1	4	0.77	0.77	0.64	0.70	0.44
2	2	0.85	0.78	0.72	0.75	0.31
3	4	0.85	0.85	0.65	0.75	0.44
4	3	0.90	0.82	0.68	0.76	0.54
5	4	0.82	0.83	0.64	0.72	0.44

Table 3.12: Intra-Scorer Consistency in Asthenia

SLT	Different scores	Pearson Corr	ICC	Kappa	Weighted Kappa	RMSE
1	5	0.69	0.68	0.49	0.57	0.44
2	2	0.91	0.90	0.75	0.84	0.31
3	3	0.91	0.91	0.75	0.84	0.38
4	1	0.91	0.90	0.83	0.87	0.22
5	9	0.59	0.45	0.25	0.36	0.86

Table 3.13: Intra-Scorer Consistency in Strain

and Cohen's Kappa. SLTs may therefore have more confidence in 'Asthenia' scoring and detecting this aspect of voice disorder than in scoring the other GRBAS components. SLT2 has 'almost perfect agreement' in 'Strain' by Cohen's Kappa and

Weighted Kappa. SLT3 and SLT4 approach 'almost perfect agreement' in 'Strain' by Cohen's Kappa and Weighted Kappa. SLT3 and SLT4 approach 'almost perfect agreement' in 'Strain' according to Weighted Kappa.

To further investigate the variation between the scores between the first and second time of scoring for each SLT the RMSE was computed. The lowest RMSE was obtained for SLT4 in 'Strain' scoring.

3.6 Conclusions so far

The tables appear to demonstrate that all five measurements, including Pearson Correlation have value as indicators of intra-scorer agreement though some interesting differences emerge. The fact that weighted Kappa is less affected by minor inconsistencies than is Cohen's Kappa probably makes it preferable. In deciding which measurement of consistency to take, the choice is probably between ICC and weighted Kappa.

It is clear that some of the scoring is rather inconsistent according to all measures. Table 3.14 presents an aggregate of the measurements of consistency over all scorers for each GRBAS component. It appears that Asthenia produces the most consistent scores whereas Roughness seems to be the most difficult component to score. Table 3.15 presents an aggregate of measurements of consistency for each scorer over all 5 GRBAS components. According to Table 3.15, Scorer 5 has the lowest level of consistency over all five GRBAS components whereas scorers 1, 2 and 4 have much higher levels of consistency. One purpose in producing these measures of consistency is to take them into account when producing a 'gold standard' for the GRBAS scoring of all the 102 subjects. This will be the subject of a later section.

Comp	Differences	Pearson Corr	ICC	Kappa	Weighted Kappa	RMSE
G	27	0.87	0.86	0.63	0.76	0.52
R	28	0.77	0.74	0.55	0.65	0.53
B	25	0.85	0.78	0.55	0.68	0.54
A	17	0.84	0.81	0.67	0.74	0.43
S	20	0.80	0.77	0.61	0.70	0.44

Table 3.14: Aggregate of measurement of consistency for each GRBAS component over all scorers

SLT	Differences	Pearson Corr	ICC	Kappa	Weighted Kappa	RMSE
1	16	0.83	0.88	0.71	0.76	0.38
2	19	0.88	0.87	0.67	0.77	0.42
3	28	0.79	0.62	0.51	0.63	0.51
4	18	0.87	0.72	0.68	0.76	0.49
5	36	0.75	0.83	0.44	0.59	0.70

Table 3.15: Aggregate of measurement of consistency for each scorer over all 5 GRBAS components

3.6.1 Inter-Scorer Consistency

'Inter-scorer consistency' is the agreement between different scorers when they score the same list of subjects. For ordinal numeric data the consistency between two scorers can be measured by Pearson correlation when the means are the same. Otherwise, ICC, Cohen's Kappa and Weighted Kappa may be employed as they were for measuring intra-scorer consistency. Cohen's Kappa treats the scores as categorical, whereas Weighted Kappa can be applied to discrete ordinal data to reflect degrees of differences between scores. Where there are more than two scorers, it is useful to have an overall measure of consistency across all scorers, and ICC readily generalises to this application. Cohen's Kappa generalises to the Fleiss Kappa which is currently applicable only to data considered categorical. A weighted Fleiss Kappa has been devised in this thesis for ordinal data such as GRBAS scoring by more than two scorers.

Table 3.16 shows the means of the GRBAS scores produced by five scorers over 102 voice samples. It may be seen that the mean is different for each GRBAS component which means that Pearson Correlation may prove unreliable as a measure of inter-scorer consistency between pairs of scorers.

SLTs	Mean G	Mean R	Mean B	Mean A	Mean S
SLT 1	0.93	0.69	0.55	0.63	0.56
SLT 2	1.2	1	0.52	0.3	0.35
SLT 3	1.14	0.57	0.68	0.76	0.84
SLT 4	1.03	0.77	0.62	0.48	0.47
SLT 5	0.94	0.61	0.33	0.38	0.51

Table 3.16: Means of GRBAS Scores

Table 3.17 presents the Cohen's Kappa and Weighted Kappa measurements obtained for each possible pair of SLT GRBAS scorers and all 102 subjects. It may be seen that there is very good agreement for ‘Grade’ between scorers 4 and 5 according to both Cohen’s Kappa and Weighted Kappa. This is interesting in view of the apparent intra-scorer inconsistency of scorer 5 in scoring Grade (see Table 3.9).

	G		R		B		A		S	
SLTs	Kappa	w-K	Kappa	w-K	Kappa	w-K	Kappa	w-K	Kappa	w-K
1,2	0.51	0.67	0.45	0.58	0.49	0.63	0.29	0.30	0.35	0.49
1,3	0.54	0.70	0.41	0.56	0.55	0.67	0.49	0.61	0.40	0.52
1,4	0.35	0.47	0.21	0.39	0.26	0.36	0.25	0.36	0.19	0.23
1,5	0.41	0.50	0.33	0.42	0.33	0.43	0.19	0.34	0.34	0.35
2,3	0.52	0.69	0.32	0.45	0.42	0.58	0.32	0.30	0.27	0.38
2,4	0.27	0.40	0.20	0.33	0.33	0.34	0.09	0.10	0.24	0.28
2,5	0.33	0.44	0.31	0.42	0.25	0.36	0.221	0.225	0.381	0.380
3,4	0.35	0.53	0.38	0.48	0.42	0.46	0.26	0.34	0.27	0.37
3,5	0.38	0.55	0.45	0.53	0.25	0.36	0.30	0.35	0.25	0.35
4,5	0.74	0.84	0.53	0.63	0.31	0.54	0.46	0.56	0.55	0.64

Table 3.17: Inter-Scorer Agreement in GRBAS scoring

According to weighted Kappa, there is substantial agreement between scorers 1 and 2 and also between scorers 1 and 3 in respect to ‘Grade’ and ‘Breathiness’. Cohen’s Kappa only considers this agreement to be moderate. Also according to weighted Kappa, scorer pairs have agreement considered less than substantial in ‘Roughness’, ‘Asthenia’ and ‘Strain’, except scorers 4 and 5 in ‘Roughness’ and ‘Strain’ and scorers 1 and 3 in ‘Asthenia’.

Table 3.18 computes the averages of the ten measurements in each column of Table 3.17 to illustrate the potential of these averages as possible alternatives to Fleiss Kappa and the newly proposed weighted Fleiss Kappa for assessing overall agreement over many scorers.

G	**G**	**R**	**R**	**B**	**B**	**A**	**A**	**S**	**S**
Kappa	w-K	Kappa	w-K	Kappa	w-K	Kappa	w-k	Kappa	w-k
0.44	0.57	0.35	0.47	0.36	0.47	0.28	0.34	0.32	0.39

Table 3.18: Kappa & weighted Kappa averaged over all SLT pairs in GRBAS scoring

Table 3.19 presents the values of traditional Fleiss Kappa and the new weighted Fleiss Kappa obtained for each GRBAS component over all 102 subjects. There is substantial difference between the two forms of Fleiss Kappa, which indicates that weighting small differences, possibly arising from marginal decisions, less than more major discrepancies in scoring may give significant improvements in the scoring.

Table 3.20 presents the ICC scoring obtained for each GRBAS component across all 5 scorers and all 102 subjects. Comparing Tables 3.19 and 3.20 it may be seen that ICC measurements are closer to the new weighted Fleiss measurements than they are to traditional Fleiss.

	Fleiss Kappa			Weighted Fleiss Kappa		
Component	P	P_e	K	P	P_e	K
G	0.6	0.28	0.44	0.63	0.13	0.57
R	0.59	0.37	0.35	0.37	0.08	0.31
B	0.64	0.43	0.37	0.52	0.006	0.51
A	0.60	0.43	0.29	0.53	0.05	0.50
S	0.61	0.43	0.31	0.52	0.02	0.51

Table 3.19: Fleiss Kappa linearly weighted Fleiss Kappa for 5 GRBAS

GRBAS	**ICC**	**Agreement**
G	0.70	Between Fair and Good
R	0.56	Between Fair and Good
B	0.57	Between Fair and Good
A	0.43	Between Fair and Good
S	0.48	Between Fair and Good

Table 3.20: ICC between the five SLTs for GRBAS

3.7 Reference GRBAS scores

The deviation of ‘reliable’ scores from a given number of trained subjective scorers taking into account inter-rater and intra-rater agreement has been considered. As mentioned earlier, we refer to these ‘reliable’ score as our ‘gold standard’ GRBAS scores. To achieve the ‘gold standard’ GRBAS scores, the following methodologies have been investigated.

3.7.1 Averages

The most obvious way of obtaining a reliable GRBAS score for subject i from n raters is to take an average. The simplest average is the arithmetic sample mean. For GRBAS scores 0,1,2,3, this gives a non-integer average score, but such scores should not cause a problem when applying machine learning. Of course, averaged scores could be rounded to the nearest integer score, but this incurs some loss of information. There are many other types of averages such as the median, mode and geometric mean. Also the arithmetic mean (and others) can be weighted to give more credibility to some scorers than others.

3.7.1.1 Median and Mode

The ‘median’ is the ‘middle’ value in a list values. Given N values, the list should first be arranged in ascending order. The median is then element $(N+1)/2$ in this ordered list if N is odd and the average of elements $N/2$ and $1+N/2$ if N is even.

The ‘mode’ of a list of values is the value that occurs most often amongst this list. Where there are two or more values that occur most and equally often, as in 0, 0, 1, 1, 2 for example, one of these values may be chosen arbitrarily, or an average may be taken.

3.7.1.2 Geometric Mean

The Geometric Mean (GM) of a list of N numbers is the Nth root of the product of the N numbers. It is often used as an alternative to the arithmetic mean and is often quite close to it especially for normal distributions. The Geometric mean can be weighted in a similar way to the arithmetic mean.

3.7.1.3 Weighted arithmetic means

Given the scores s1, s2, s3, s4, s5 from five scorers, a weighted arithmetic mean is:

$$W_{\mathrm{m}} = \frac{w_1s_1+w_2s_2+w_3s_3+w_4s_4+w_5s_5}{w_1+w_2+w_3+w_4+w_5} \tag{3.27}$$

where the w values are the weighting factors. Choosing w_1=3 with all others equal to 1 would weight the contribution of scorer 1 as three times that of the other scorers. Scorer 1 would be considered more reliable (for some reason) than the others.

The weighted average is actually the normal average obtained from a supposed larger group of scorers than we actually have, where the results of each scorer is replicated a number of times. In the example above with w_1=3 and $\mathrm{w}_2 = \mathrm{w}_3 = \mathrm{w}_4 = \mathrm{w}_5 = 1$, it is considered that there are 8 scorers with the score from scorer1 repeated three times.

The number of times a scorer is replicated may be done according to the assumed ‘reliability’ of the scorer. Scorers with high reliability are replicated more times than those with lower reliability. The weights are measures of reliability or ‘beliefs’ about the reliability based on some evidence. Each weight could be interpreted as the probability that the scorer would be correct. They may be also considered as measure of confidence. So if a scorer has a reliability weight 1, we have total confidence in his scoring. If he has weight of zero we have no confidence in his scoring. We can think of these weights as Bayesian measures of the probability of being correct, or 'confidence'.

Define an n by n 'Reliability Matrix' (RM) for each GRBAS component where n is the number of scorers. In our application n is equal to 5. There are five such matrices, R_G, R_R, R_B, R_A and R_S each with non-diagonal entries derived from the 'weighted Kappa' columns of Table 3.17. The diagonal entries reflect intra-scorer agreement for each scorer and are derived for 'Grade' from the 'weighted Kappa' columns of Table 3.9. The diagonal entries for R_R, R_B, R_A and R_S are similarly derived from Tables 3.10, 3.11, 3.12 and 3.13. We have therefore decided to adopt weighted Kappa as our measure of both intra-scorer and inter-scorer consistency, bearing in mind that the average of all pair-wise weighted Kappa measurements is very close to the new weighted Fleiss Kappa proposed in this thesis. The RM matrix obtained for 'Grade' is shown in Equation (3.28). Equation (3.28) shows the 'RM' for Grade where the entries on the main diagonal were obtained by intra-scorer consistency computations and the entries outside of the main diagonal were obtained by inter-scorer consistency computations.

$$R_{\mathrm{G}} = \begin{bmatrix} 0.88 & 0.67 & 0.7 & 0.47 & 0.5 \\ 0.67 & 0.78 & 0.69 & 0.4 & 0.44 \\ 0.7 & 0.69 & 0.71 & 0.53 & 0.55 \\ 0.47 & 0.4 & 0.53 & 0.82 & 0.84 \\ 0.5 & 0.44 & 0.55 & 0.84 & 0.62 \end{bmatrix} \tag{3.28}$$

To make use of this matrix to obtain a 'gold standard' score for grade for subject i, we proceed as follows: First we express the 'pair-wise' weighted average between the grade scores given to patient i by scorers L and M as follows where $s_L(i)$ and $s_M(i)$ are the Grade scores and w_L and w_M are weights read from the diagonal entries of the reliability matrix R_G.

$$P_{\mathrm{LM}}(i) = \frac{w_{\mathrm{L}} \times s_{\mathrm{L}}(i) + w_{\mathrm{M}} \times s_{\mathrm{M}}(i)}{w_{\mathrm{M}} + w_{\mathrm{L}}} \tag{3.29}$$

Therefore, for Grade :

$$P_{\mathrm{LM}}(i) = \frac{R_{\mathrm{G}}(L,L) \times s_{\mathrm{L}}(i) + R_{\mathrm{G}}(M,M) \times s_{\mathrm{M}}(i)}{R_{\mathrm{G}}(L,L) + R_{\mathrm{G}}(M,M)} \tag{3.30}$$

The gold standard score,$G_{GS}(i)$, for Grade for patient i is then expressed as the weighted average of $P_{LM}(i)$ over all possible scorer pairs L,M where the weighting for each pair is read from the appropriate non-diagonal entry of $\mathrm{R_G}$. Therefore:

$$G_{\mathrm{GS}}(i) = \frac{\sum_{L=1}^{c-1} \sum_{M=2}^{c} R_{\mathrm{G}}(L,M) P_{\mathrm{LM}}(i)}{\sum_{L=1}^{c-1} \sum_{M=2}^{c} R_{\mathrm{G}}(L,M)} \tag{3.31}$$

The first step weights each score by the ‘consistency’ of the scorer and the second step weights each paired component of the average by the inter-scorer consistency of the pair of scorers. The procedure is repeated for R, B, A and S to obtain a ‘gold standard score’ for each GRBAS component for each subject.

3.8 ‘Gold standard’ GRBAS reference scores

Appendix B presents a table of the reference scores obtained by applying equations (3.28) to (3.31) for ‘Grade’, for each subject i in the range 1 to 102. The corresponding reference scores for ‘R’, ‘B’, ‘A’, and ‘S’ were similarly calculated for this table. The table also presents the unweighted averages of the scores for each subject (avG, avR, avB, avA, avS). Figures 3.2, 3.3 3.4, 3.5 and 3.6 show a ‘histogram’ frequency distribution of the differences between the unweighted averages of the scorers and gold standard for ‘Grade’, ‘Roughness’, ‘Breathiness, ‘Asthenia and ‘Strain’ respectively.

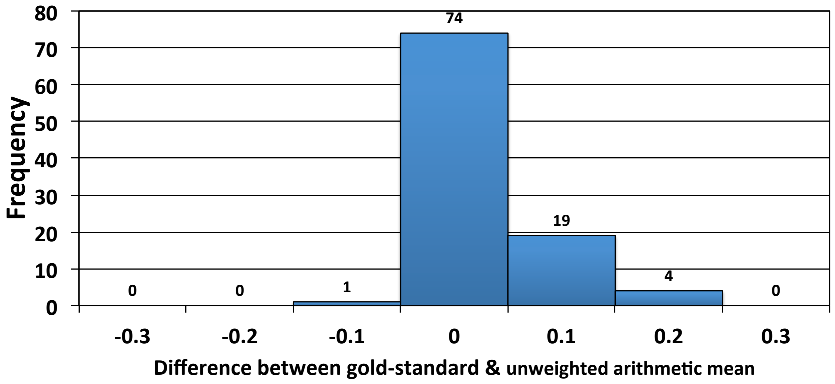


Figure 3.2: Histogram of differences between gold-standard and unweighted arithmetic mean for Grade as listed in Appendix B. Both rounded to one decimal place. The frequency is the number of subject out of the total of 102

The maximum difference is 0.3 that is obtained for Grade and Roughness with frequency 74. The minimum difference is -0.3 is achieved for strain with frequency

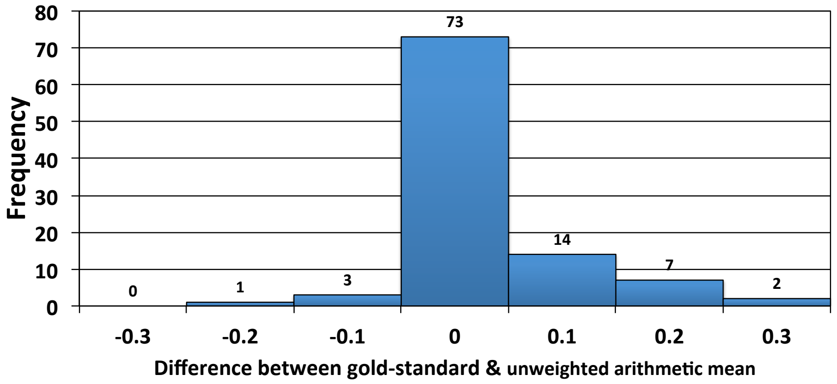


Figure 3.3: Histogram of differences between gold-standard and unweighted arithmetic mean for Roughness as listed in Appendix B. Both rounded to one decimal place. The frequency is the number of subject out of the total of 102

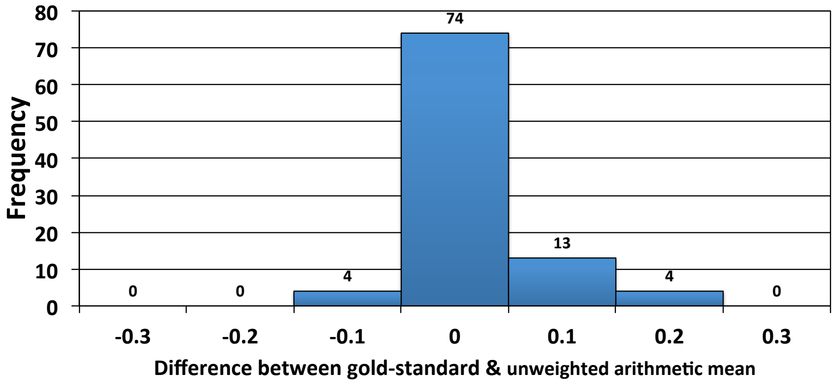


Figure 3.4: Histogram of differences between gold-standard and unweighted arithmetic mean for Breathiness as listed in Appendix B. Both rounded to one decimal place. The frequency is the number of subject out of the total of 102

1. The RMSE values between the unweighted averages of the scorers and gold standard for Grade, Roughness, Breathinees, Asthenia and Strain are 0.06, 0.08, 0.06, 0.07 and 0.07 respectively. The variation between the unweighted averages and gold standard are small. With less reliable scorers, the differences may have been greater. The‘reliable’ scores for training the machine learning algorithm in Chapter 5 is computed with straightforward average of all five scorers, since the final results of Chapter 3 were not available when these tests were run. For most of this work, only 3 scorers were available, though we were able to update the averages when the extra two scorers became available.

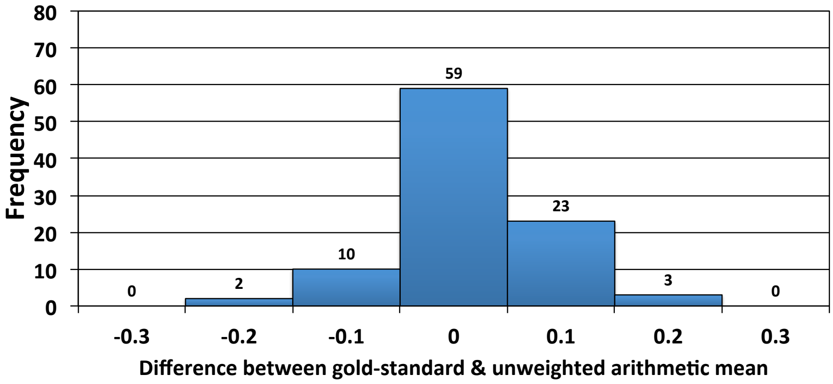


Figure 3.5: Histogram of Differences between gold-standard and unweighted arithmetic mean for Asthenia as listed in Appendix B. Both rounded to one decimal place. The frequency is the number of subject out of the total of 102

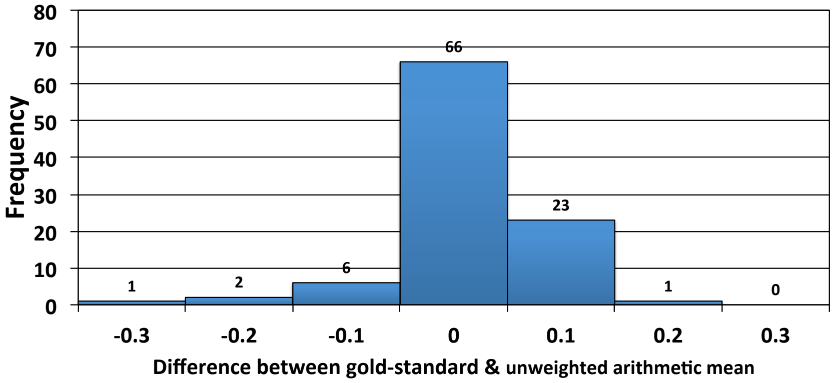


Figure 3.6: Histogram of Differences between gold-standard and unweighted arithmetic mean for Strain as listed in Appendix B both rounded to one decimal place. The frequency is the number of subject out of the total of 102

3.9 Conclusions

A data-base of voice recordings from ‘patients’ and ‘control’ participants has been provided by MRI Hospital, and a mechanism for providing clinicians with a convenient means of applying traditional subjective GRBAS scoring techniques to these recordings has been described in this Chapter. Methods of evaluating the ‘intra’ scorer consistency of individual scorers and ‘inter’ scorer consistency between different scorers are explored and have been applied to the data-base. The data-base, the scorings and the scorer evaluations are to be used as the means of developing objective GRBAS scoring by computer. In order to do this, we need to determine a ‘reliable’ (‘gold standard’) set of GRBAS scores using the scores obtained from five scorers and measurements of the intra-scorer and inter-scorer consistency of the scores. There was found to be considerable inconsistency in some of the scoring and also considerable variability from scorer to scorer. Subjective ratings can be confounded by factors such as the listener's experience, the listeners perceptual bias, the type of rating scale used, the listener's fatigue, the perceptual sensitivity of the listener to particular voice features and to the voice sample being evaluated, and many other factors. Investigating existing and new ways of measuring consistency, for example using correlation and various forms of Cohen Kappa including a new weighted Fleiss Kappa has led us to address the problem of how to make the best of the data provided, with its inherent inconsistencies. An approach to this problem has been presented using averaging weighted by measures of consistency. Whether this is the best possible approach has not been established, and there are clearly many possible solutions to the problem.

It was found that SLT scorers showed better consistency in scoring some aspects of GRBAS than others. They seemed more confident in scoring Asthenia than the others GRBAS components, and this finding was confirmed when either the Cohen's Kappa or the weighted Cohen Kappa was used as the measure of intra-scorer consistency. For GRBAS, there are good reasons for preferring the weighted version of Cohen's Kappa since it reflects the magnitudes of scoring differences. Minor differences perhaps resulting from marginal decisions are de-emphasized which seems a good thing to do. The weighted Kappa showed intra-scorer consistency to be 'almost perfect' and 'substantial' in the Grade and Asthenia components respectively. On the other hand, Kappa, Weighted Kappa and Fleiss' Kappa measurements indicate the inter-scorer agreement between SLT scorers is mostly moderate. 'Gold standard' reference scores, taking into account intra-scorer and inter-scorer consistency, have been produced and may be compared with unweighted average scores. Both sets of scores are available for use with subsequent model training. The differences between these two sets of scores are small but significant. It has been claimed [Bla] that the assessment of observer agreement is one of the most difficult areas in the study of clinical measurement. There is much further work that could be usefully done in this field.

Chapter 4

Feature Measurement

4.1 Introduction

This chapter is concerned with the digital signal processing (DSP) techniques that are used to measure characteristic features of voice recordings. There are many reasons for wishing to measure such features. Speech processing is a very mature topic which has been developed over many years for speech recognition and synthesis, coding and compression for mobile telephony, speaker recognition, noise elimination and many other applications. It has also been widely applied in the medical field for assessing voice problems, detecting emotional states, teaching deaf children to speak and many other activities.

This thesis is concerned with voice quality assessment; in particular the objective measurement of GRBAS scores. There are many characteristic features of voice that may be observed from the acoustic waveform that reaches the ear of a listener. Published and commercialised DSP software already exists for measuring voice features. Available software uses a variety of different methods. Even when measuring the same phenomena, these methods often produce different results. It is therefore necessary to understand the DSP techniques that are traditionally used for feature measurement in order to decide which techniques are appropriate and likely to be reliable. In some cases it is instructive to reproduce and evaluate well known techniques which are often very simple, such as the formulae for jitter and shimmer mentioned in Chapter 2. In other cases, techniques whose precise details are not disclosed may be considered. For example the Cepstral/Spectral Index of Dysphonia (CSID) in the ADSV software suite (see Chapter 2) may provide useful measurements even though its algorithm is a commercial product and not fully disclosed. In either case, it is important to understand the reliability and limitations of the results produced. With this knowledge, it is often possible to improve the performance of well known DSP techniques by adapting them to the application in question. At least it may be possible to understand why they fail when they do fail. Apart from CSID, many published or commercialised techniques are applicable only to 'sustained' vowels whereas one of the aims of this thesis is to measure features also from 'connected' or natural spoken sentences. The purpose of this chapter is to survey and examine the features that may be usefully measured and to investigate the DSP mechanisms that are or may be employed to do the measurement. In this chapter large N will be used to denote the length of analysis frame in sample. L will be used by the cross correlation method to denote the length of abutting sub-frames within each frame of N-samples. As will be seen in section 4.3.5 the basis of the cross correlation technique is to search for the value of L that optimises the similarity between the abutting segments.

4.2 Some Background

4.2.1 Recording and digitising Speech

Capturing segments of speech via a microphone converts the variation in air-pressure that constitutes sound into the variation of a voltage. This voltage may be sampled and then digitised so that segments of voice may be stored in a computer. Engineering constraints must be imposed on the sampling rate and the number of bits used to represent each sample to make sure that the sound is represented with appropriate fidelity. But these constraints are easily satisfied by modern equipment. Once correctly digitised and stored in computer files, the speech can be played out repetitively and also displayed visually as a graph of air-pressure against time. Such a graph is referred to as a waveform.

There is much to be learned simply by observing speech waveforms. Since speech and music may be perceived by humans over a bandwidth from about 50 Hz to 20 kHz, a standard for 'high fidelity' sound was proposed long ago. With the advent of compact disk (CD) recording, this led to a very widely used digitisation standard often referred to as 'CD quality' sound. This standard requires speech to be band-limited to the frequency range 0 to 20000 Hz and sampled at 44100 Hz with 16 bits allocated to represent each sample. Stereo sound requires two simultaneous samples of sound to be captured at 44,100 Hz. This sampling frequency is widely used for music, but is at variance with current practice in voice telephony where speech is restricted in bandwidth to the frequency range 300 to 3400 Hz and sampled at 8 kHz. Speech is largely intelligible at this lower bandwidth and sampling rate, and is even amenable to further compression of the required bit-rate without severe loss of quality or intelligibility. However, telephone quality speech is not well suited to the application we are concerned with because there is a quality loss that may not be noticed by telephone users, but which may be very important in assessing true voice quality. Therefore standard 'CD quality' sound will be used throughout this thesis meaning that the bit-rate will be $44100 \times 16 = 705600$ bits per second. Uniform quantisation without compression will be used, and steps will be taken to ensure that a reasonable signal-to-quantisation-noise ratio is achieved. Speech segments can thus be conveniently stored in 'wav' files which are an industry standard and well supported by commercial and academic software.

4.2.2 Voiced and unvoiced speech

There are characteristic features of speech that can be seen immediately in waveforms displayed on a computer screen. For example one can observe the difference between 'voiced' speech which is pseudo-periodic and 'unvoiced' speech which has no periodicity. The cause of this difference is the human speech production mechanism which can produce sound from the vocal cords or from turbulent flow created at some constriction within the vocal tract. Voiced speech, created by pseudo-periodically vibrating vocal cords, produces vowels while unvoiced speech created by turbulent flow produces the sound heard within consonants. Voiced speech can remain approximately stationary (non-changing) for periods of time approaching or exceeding one tenth of a second, whereas unvoiced speech tends to be transitory and fast changing, maybe lasting only 0.01 seconds or less. Distinguishing voiced from unvoiced speech is a significant and important challenge, and not always easy to do reliably even for normal talkers. Doing this for quality impaired voices is even more difficult. The challenge with voiced/unvoiced detection is detecting periodicity as will be discussed later.

4.2.3 Analysing voiced speech

For regions of the speech waveform identified as voiced, the next challenge is to measure the periodicity and the nature of this periodicity. The voice will have a fundamental frequency which is determined by the rate of vibration of the vocal cords. This determines the pitch of the voice which is normally constantly changing. Measuring the fundamental frequency can sometimes be difficult for normal talkers though it is required for many speech bit-rate compression techniques used in mobile telephony and Voice over IP (VoIP). Doing this for quality impaired voiced can be very difficult. Therefore voice quality assessment techniques employed in the medical field try not to be too critically dependent on estimations of fundamental frequency. Where a fundamental frequency is discernible, it will normally be changing with time. Some of this variation will be due to natural intonation; a gradual lowering or raising of the pitch of the voice when answering or asking questions, for example. But there may also be a much more rapid variation in fundamental frequency which is referred to as 'jitter'. Jitter is a form of frequency modulation that is so rapid that it is perceived as roughness in the sound of voiced speech. A related parameter is 'shimmer' which is the effect of rapid pitch-cycle to pitch-cycle variation in the amplitude of the speech signal. Shimmer is a form of rapid amplitude variation which is again perceived as roughness. Jitter and shimmer are often an indication of abnormality in the way the vocal cords are vibrating during voiced speech.

4.2.4 Sound due to turbulent air-flow

Most sound energy during voiced speech is produced by the vocal cords 'snapping' closed after opening more slowly. This mechanism relies on a natural elasticity and uniformity of the tissue which may be affected by infection, inflammation or other damage. Turbulent air-flow which is responsible for unvoiced speech creates a non-periodic waveform of wide bandwidth which can extend to the upper limit of the 20 kHz bandwidth of 'CD quality' speech. In contrast, the energy of voiced speech tends to fall off rapidly beyond about 4 kHz.

Because turbulent air-flow is a chaotic process causing random collisions among molecules of air, it is well modeled as a random process. Using a random number generator to create voltage samples can produce waveforms that resemble and sound like unvoiced speech. Such waveforms are sometimes referred to as 'noise-like' waveforms because certain types of unwanted noise, for example wind noise in a car, the background sound from a badly tuned radio or television and quantisation noise, tend to sound like this. The frequency spectrum of such 'noise' is wide and is often close to being flat or 'white' meaning that all frequencies within the audible range are equally present.

Turbulent air-flow can be present within voiced speech. It is a characteristic of the way some people, especially females, speak naturally, and is often referred to as 'breathy' speech. But it can also be an indication of voice pathology where, for some reason, the vocal cords are not closing completely during their so called 'closed glottis' phase. In such cases, the voiced speech may not be strongly periodic because of the presence of some turbulent air-flow causing the introduction of a 'noise-like' supplement. The 'noise-like' random signal is often considered to be added to the pseudo-periodic voiced component. Aperiodicity during voiced speech is possibly the strongest indication of voice pathology and it is often caused by 'breathinesss' as just described. When it is, the situation can be quantified by a 'harmonic-to-noise' ratio where the harmonic part of the speech is that part considered to be pseudo-periodic. Even when the causes of aperiodicy are more complex than the addition of sound due to turbulent flow, the degree of aperiodicity may still be well represented by a 'harmonic to noise ratio', considering the source of aperiodicity to be modeled by an additive random component. Detecting the presence of noise-like features has proven to be reliable for detecting voice disorders, since most pathological voices present some degree of aperiodicity. Such aperiodicity may be quantified by measurements of 'harmonic to noise ratio' (HNR) [YGB82], jitter, and shimmer [WFC95]. However, these measurements have been shown to be unreliable predictors of dysphonia in many studies [HH96, HAMB^{+}03].

One reason for this unreliability may be that what one perceives as dysphonia may not be logically associated with any one perturbation measure. Another possibility is that some of these measurements have relied on the ability to determine the fundamental frequency F_0. Small errors in determining F_0 can lead to significant errors in measuring perturbations. Because of the difficulty in determining F_0, accurate measurements of periodicity in dysphonic voice samples that are marginally aperiodic are often difficult to obtain. Measures of cross-correlation and Cepstral Peak Prominence (CPP) attempt to solve this problem [HH96, HAMB^{+}03]. As the degree of voice abnormality increases more noise-like characteristics appear and replace the harmonic structure. Therefore the degree of voice abnormality can be evaluated by judging the extent to which noise-like features replace the harmonic structure in the sustained vowel. HNR, Jitter and Shimmer attempt to quantify features that destroy the harmonic structure of voiced speech.

The ADSV approach uses Cepstral Coefficients to characterize the aperiodicity of pathological voices in a different way [AR06, ARJ^{+}10]. In this thesis, some algorithms have been developed in MATLAB for quantifying voice characteristics, and other algorithms within the research literature and within the Kay-Pentax commercial software tools have been adopted. The commercial ADSV (Analysis of Dysphonia in Speech and Voice) tool provided by Kay-Pentax has been researched for this purpose.

4.3 DSP techniques for measuring voice features

Important aspects of speech processing in many applications are detecting periodicity, distinguishing voiced from unvoiced speech and measuring the fundamental frequency of voiced speech. These aspects are used in speech coding for mobile telephony and voice over IP and a great deal of work has been invested in them over the years [RS11]. They are also important for other speech processing applications, especially in the medical field, and are a vital ingredient of the subject area of this thesis. We have investigated DSP algorithms for measuring 'pitch frequency', 'rapid pitch frequency variation (jitter)', 'rapid amplitude variation (shimmer)', 'energy', 'low to high spectral energy', 'degree of periodicity' and 'harmonic to noise ratio'. Leaving aside some very rudimentary time-domain approaches such as zero-crossing counting, filtering and phase-locked loops, the most common approaches can be divided into three classes: FFT frequency-domain methods, cepstral methods and autocorrelation function methods. The cross-correlation method presented later, may be considered a special case of autocorrelation function methods, though with subtle and important differences.

4.3.1 Fast Fourier Transform (FFT) methods

Discrete Fourier Transform (DFT) of a segment $\{x[n]\}_{0,N-1}$ of a real or complex-valued digital signal is given by equation (4.1) where x[n] is the nth sample of the time-domain segment. The result is a frequency-domain representation of the original segment in the form of a complex-valued segment $\{X[k]\}_{0,N-1}$ where X[k] is the kth frequency domain coefficient representing DFT spectral content at frequency $F = k \times (F_S/N)$ Hz where F_S is the assumed sampling frequency in Hz. Equation (4.2) is the 'inverse DFT' which reverses the effect of the DFT to generate the time-domain sequence $\{x[n]\}_{0,N-1}$ from the frequency-domain sequence $\{X[k]\}_{0,N-1}$.

$$X[k] = \sum_{k=0}^{N-1} x[n] e^{-i\frac{2\pi}{N}kn} \quad k = 0, 1, 2, ..., N-1 \tag{4.1}$$

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{i\frac{2\pi}{N}kn} \quad n = 0,1,2,....,N-1 \tag{4.2}$$

The FFT is a computationally efficient method for computing the DFT of a segment of a digitized signal. It produces exactly the same result as evaluating the DFT formula directly, but much faster. The number of arithmetic operations, multiplications and additions, for the direct computation of the DFT is approximately proportional to N^2 and $N(N-1)$ respectively, but for the FFT algorithm this reduces to approximately $(N/2)\times\log 2N$ and $N\times\log 2N$. This difference becomes dramatic when N has a large value. An ‘inverse FFT’ exists as a similarly fast and efficient way of implementing an inverse-DFT.

4.3.2 Short term Fourier transform (STFT)

The Short-Time Fourier Transform (STFT), is a way of analysing non-stationary signals, whose statistical characteristics vary with time. Speech is dynamic in that it is time-varying. Although the characteristics of voiced speech within a spoken sentence may stay fairly constant for many pitch-periods during a ‘sustained vowel’, maybe up to 200 ms, they will not stay exactly the same. Unvoiced segments of speech tend to be more short-lived and change more rapidly.

The STFT is normally applied pitch-asynchronously with fixed-length analysis segments. However, it can be applied pitch-synchronously using variable length segments with the length chosen to contain an exact number of pitch-cycles. Clearly, the pitch-period must be computed before applying pitch-synchronous Fourier analysis, and this can be difficult with pathological voices. Variations in the fundamental frequency and amplitude make this even more difficult and can negate the advantages of pitch-synchronous analysis. Despite these difficulties, this is how the MDVP software appears to calculate ‘harmonics to noise ratio’ (HNR or NHR).

Pitch-asynchronous STFTs analyse segments that are short enough to be considered approximately stationary. Speech segments of length typically around 20 ms are often considered short enough for approximate stationarity, though segments as short as 10 ms and as long as 50 ms are sometimes used. The terms ‘window’ and ‘windowing’ are commonly used to describe the extraction of a speech segment by multiplying by a time-domain ‘window’ function {w[n]} which becomes non-zero only for the short time-span of the STFT. A succession of STFT computations may be performed with windows that move forward in time. The result is a series of STFT spectra that constitutes a ‘spectrogram’. If the time-span of the window is sufficiently narrow, each frame extracted can be viewed as approximately stationary so that the DFT can be used. If each segment extracted is of length N samples and its samples are re-indexed to start at zero, the STFT may be defined as in Equation (4.3).

$$X[k] = \sum_{n=0}^{N-1} w[n] \times x[n] \times e^{\frac{-2\pi ikn}{N}} \quad k = 0, 1, ..., N-1 \tag{4.3}$$

The windowing function {w[n]} is described as rectangular if its value is constant within its non-zero region. However, non-rectangular windows which emphasize the central part of the extracted segment, at the expense of samples close to the edges, are more commonly used. The Hamming window is perhaps the most widely used form of non-rectangular window. It may be shown that the Hamming window reduces spectral spreading [OSB$^+$89] caused by a pitch-asynchronous fixed-length segment extraction process. Two other devices are commonly used when applying the STFT: zero-padding and decimation. Zero-padding increases the number of frequency-domain samples by simply inserting extra zero valued time-domain samples. The FFT gives N samples in the frequency-domain when there are N time-domain samples, and increasing N often makes the FFT spectrum easier to visualize and process.

Decimation reduces the bandwidth and sampling rate of a signal which is useful when the STFT analysis is to be focused at the lower frequency end of the spectrum. In this work, real speech is sampled at 44100 Hz which means that the STFT will produce spectra over the frequency range 0 to 22050 Hz. However the energy in voiced speech tends to be concentrated below about 4 kHz, so it is often useful to decimate the signal to a sampling rate of about 8 kHz before performing the STFT, when we are concentrating on voiced speech only. Such decimation involves digital filtering to reduce the bandwidth to around 0 to 4 kHz, and then re-sampling at about 8 kHz. A common approach down-samples the speech by a factor 5.5, which is achieved by up-sampling to a sampling rate of $2 \times 44100 = 88200$ Hz before down-sampling by a factor 11. The resulting sampling rate of 8018 Hz is close enough to 8 kHz for the work in this thesis.

To convert a recording of speech to the frequency-domain for subsequent analysis, the following steps may be carried out:

1. Apply decimation to achieve the required band-width and sampling rate

2. Divide the decimated audio signal into segments of suitable length.

3. With pitch-asynchronous methods, multiply each segment with a Hamming window to reduce spectral spreading.

4. Apply zero-padding to each segment to achieve the required spectral resolution.

5. Compute the FFT of each modified segment.

It is common to examine the magnitudes of the complex-valued spectra thus produced. From this, the 'spectral envelope' is discernible as a smoothed curve joining all the magnitude spectral samples.

Any periodicity in the time-domain signal may now be discernible as spectral 'lines' or peaks in the frequency-domain at samples corresponding to the fundamental frequency and its harmonics. Such peaks will usually be clear and obvious for normal voices and the fundamental frequency of a frame of voiced speech can therefore be measured from such a magnitude spectrum. There will usually be some spectral spreading which increases around the higher order harmonics due to the segment being not precisely stationary. There may also be some spectral content around the harmonics due to 'noise-like' effects including 'breathiness', and the effects of jitter and shimmer.

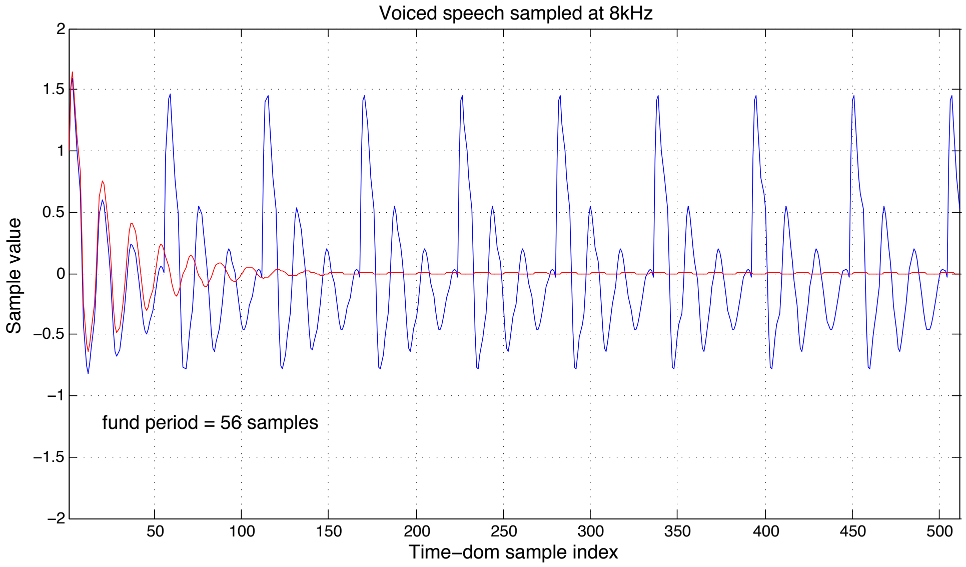
For dysphonic voice samples, finding the fundamental frequency from the STFT magnitude spectrum may not be an easy task to perform, because the amount of 'noise' might be much greater and the effects of 'jitter' and 'shimmer' may be much greater than with normal voices. Distinguishing voiced from unvoiced speech is possible in the pitch-asynchronous STFT frequency-domain, but not so easy, even for normal voices, when there are significant changes within the frame, as occur at transitions between vowels and consonants. Measuring the 'degree of periodicity' is also not very easy in this domain and producing a 'harmonic to noise' ratio can be very difficult during transitions.

The FFT is essentially an averaging process over the period of time defined by the duration of the window. Each FFT frequency-domain sample is an average of the instantaneous spectral energy over the whole frame of typically 20 ms. The pitch-asynchronous FFT is useful for measuring average spectral characteristics, such as 'spectral tilt' or the ratio of energy in different spectral bands. Spectral tilt measures the degree to which the short-term spectrum is biased towards the low frequency or high frequency extremes of the frequency band being analysed. However, significant changes can occur even over a time-window as short as 20 ms and it is often just at times when such changes occur that we most need an estimation of the voicing and periodicity. Therefore, the pitch-asynchronous FFT has limited use in high accuracy voice parameter estimation, and most modern techniques tend to operate in the time-domain using autocorrelation and cross-correlation based techniques. Some parameter estimation techniques use the pitch-asynchronous FFT, because of its computational efficiency, to obtain initial rough estimates of parameters which are later refined by autocorrelation domain techniques [Rab77]. As mentioned earlier, some techniques use pitch-synchronous FFT techniques, but these rely on accurate estimations of fundamental frequency which may be difficult to obtain.

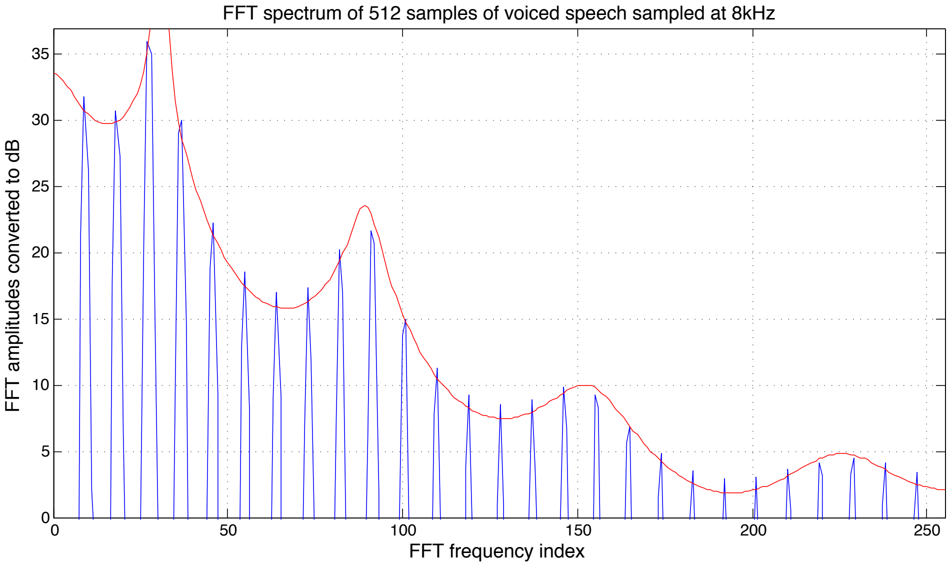
4.3.3 Cepstrum methods

The cepstrum is the inverse Fourier transform of the logarithm of the Fourier transform of a speech segment. In digital signal processing, the Fourier transform and its inverse are invariably implemented by the FFT and its inverse. Typically the FFT is applied to fixed-length speech frames of duration about 20 to 50 ms with windowing, decimation and zero-padding applied to achieve the required resolution and to reduce artifacts such as spectral spreading. Considering the speech spectrum to be the convolution of a periodic excitation e(t) of fundamental frequency less than about 200 Hz, and the impulse-response h(t) of a filter modeling the resonances of the vocal tract, as seen in the frequency-domain this becomes the product of two spectra $E(f)\ and\ H(f)$. Applying the logarithm to this product produces: $log(E(f)) + log(H(f))$ which can be computed in complex logarithms, or purely in magnitude spectral terms. In practice, we usually convert the spectrum to un-normalized deciBels (dBs) by taking $20 \times log10$ of the FFT magnitude spectrum. The effect of $E(f)$ for voiced speech is a series of spectral lines or peaks at the fundamental frequency and its harmonics. The effect of $H(f)$ is to create the 'spectral envelope' of the magnitude spectrum determining the heights of the peaks. These effects are illustrated in Figure 4.1b for the 64 ms segment of synthesised voiced speech shown in Figure 4.1a.

The synthetic speech was generated by exciting a ninth order all-pole digital filter by a periodic series of discrete time impulses. The sampling frequency, F_S, is 8 kHz and the duration is 64 ms which is a little longer than we normally use with real connected speech. The pole pairs were at 480 Hz, 1400 Hz, 2400 Hz and 3520 Hz with



(a) Time domain.



(b) Frequency domain.

Figure 4.1: Synthetic vowel in time domain (a) and frequency domain (b). The time-domain sample index refers to the speech sampled at 8 kHz. The FFT frequency index refers to the 512 point FFT of the synthetic vowel. The frequency represented by k is $F_s \times k/512$ where F_s= 8kHz. The red graph is the impulse-response.

amplitudes 0.97, 0.93, 0.85 and 0.8 respectively. There is a fifth pole which is real with amplitude 0.9 to model the combined effect of lip-radiation and the assumed shape of the glottal excitation pulse produced by the vocal cords. This hypothetical example of speech is highly resonant at the first pole and corresponds (roughly) with the vowel /a/. The impulse-response of the all-pole digital filter is shown in red superimposed on the speech in Figure 4.1a.

The real cepstrum is obtained by computing the inverse FFT of the log spectrum shown in Figure 4.1b. The IFFT is linear, therefore the result is the sum of the inverse FFT of $20 \times log10(E(f))$ and the inverse FFT of $20 \times log10(H(f))$. These two terms are now separable because of their frequency content. The cepstrum thus obtained is shown in Figure 4.2. As seen in Figure 4.1a, the impulse-response of the all-pole filter, which is the IFFT of $H(f)$, decays to very low values within about 50 samples. The IFFT of $log(E(f))$ will die away even faster as can be seen as the red graph superimposed on the cepstrum shown in Figure 4.2. The remaining part of the cepstrum is due to $E(f)$ which corresponds to the vocal tract excitation signal $e(t)$. The periodicity of $e(t)$ is now easily seen in the cepstrum, especially for the artificial speech example where it remains exactly constant and there is no additive noise.

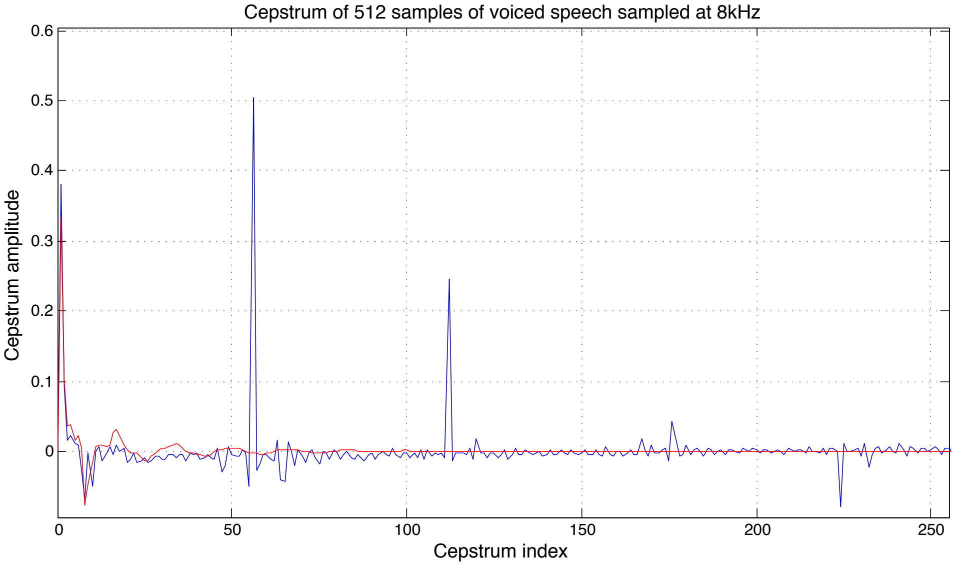
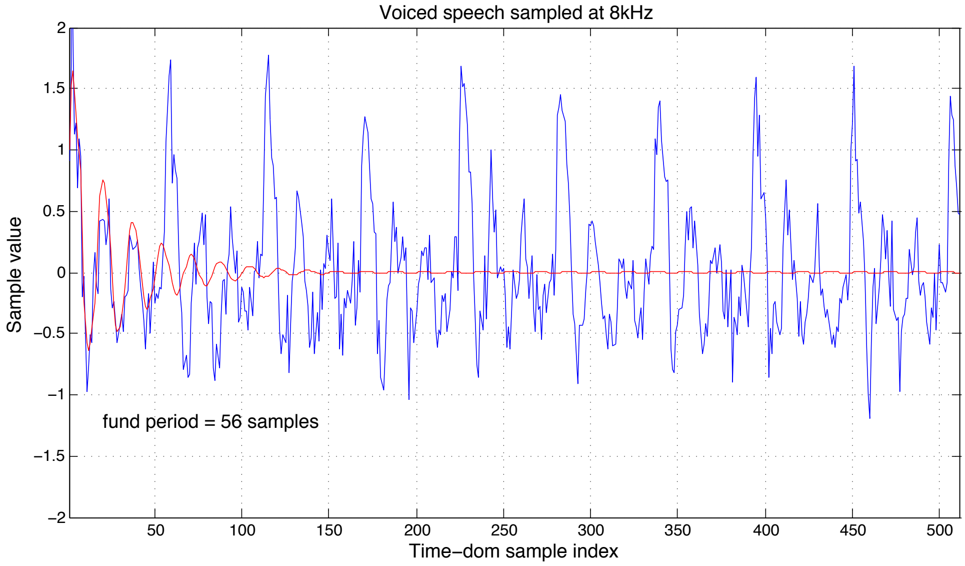
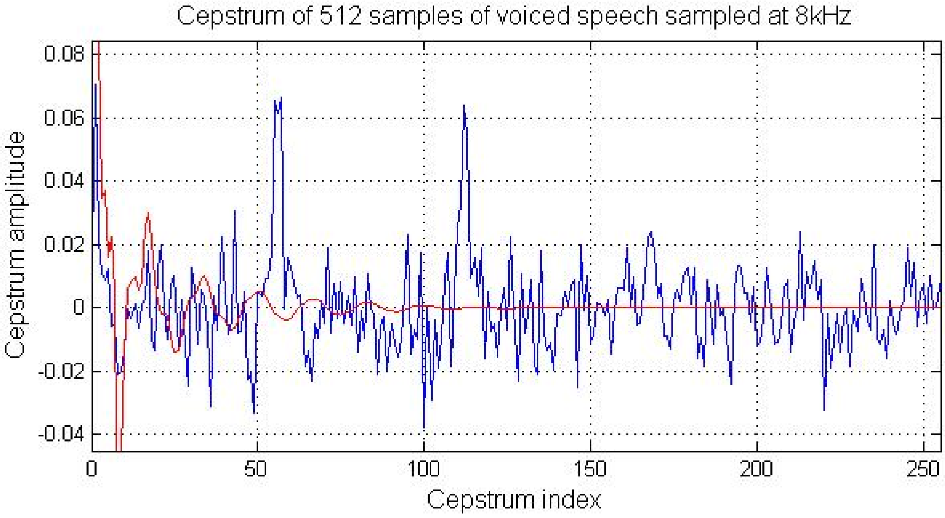


Figure 4.2: Real cepstrum for artificial speech. The Cepstrum index k represents time in units of 512/ F_s= 64 ms. Therefore k represents (64k) ms in time (or quefrequency). The red graph is the impulse-response.

The fundamental frequency of $e(t)$ and its harmonics create the cepstral lines or peaks seen in Figure 4.2. In this case the fundamental frequency is unmistakable as $Fs/56$ where 56 is the cepstrum index (or ‘quefrency’) of the first cepstral peak and F_S = 8 kHz. This peak is very ‘prominent’. The fundamental period is 56 samples, as we know already. The cepstrum is therefore useful for determining the fundamental frequency of speech, but it is also widely used for estimating the ‘degree of periodicity’ when the speech is not exactly periodic, perhaps because of additive noise, jitter, shimmer or other effects. Figure 4.3a shows the same 20ms segment of synthetic speech as in Figure 4.1a, but now with additive white Gaussian noise of zero mean, variance 0.2 and band-limited to 4 kHz.



(a) With added noise.



(b) Cepstrum for figure 4.3a.

Figure 4.3: Synthetic vowel with noise time domain (a) and cepstrum (b). The time-domain sample index refers to the speech sampled at 8 kHz. The Cepstrum index k represents time in units of 512/ F_S= 64 ms. Therefore k represents (64k) ms in time (or quefrequency). The red graph is the impulse-response.

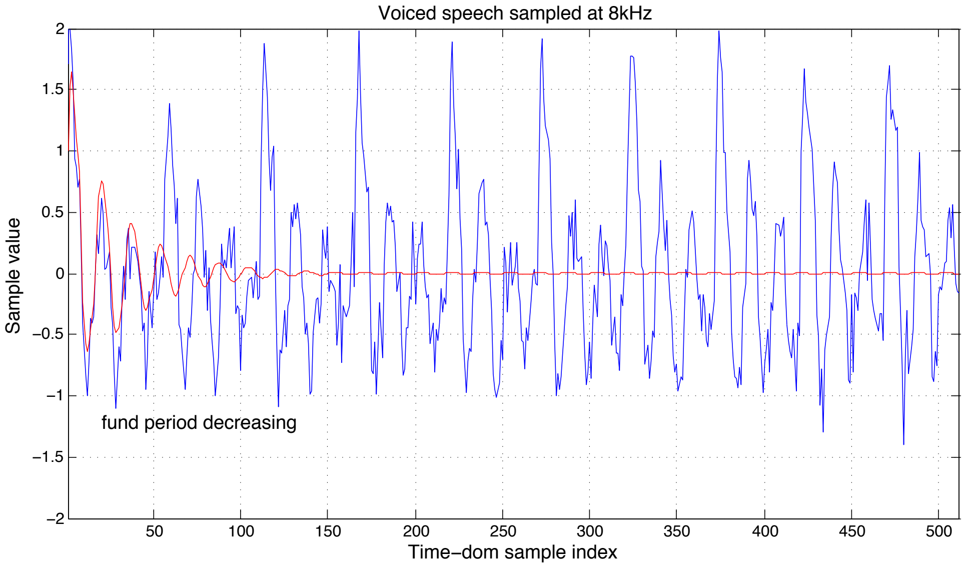
The fundamental frequency is still easily discernible as the noise is not very severe. However in Figure 4.3b, the cepstral peak corresponding to the fundamental frequency is not so 'prominent'.

The effect of the added noise is seen in the cepstrum, and the 'cepstral peak prominence' (CPP) is, loosely speaking, the degree to which the peak is prominent above the contribution of the noise. It is intuitive that a highly periodic signal should have a more prominent cepstral peak than a less periodic signal. However, while the meaning of periodic is clear, the concept of 'degrees of periodicity' is not often discussed in the literature and has not been addressed so far in this thesis. It will be later. Disregarding this issue for the moment, it is considered [HH96] that what is needed is a measure of the prominence of the peak rather than its absolute amplitude. Several methods of quantifying CPP have been proposed in the literature [HH96, AR06, ARJ$^+$10]. The method adopted by many researchers [HH96, AR06, ARJ$^+$10] is to fit a linear regression line to the samples of the cepstrum. This is a line for which the sum of the squared distances to the straight line from all points in the graph is minimised. There is an argument for leaving out the cepstral peaks from this calculation so that only the effects of the noise or other causes of aperiodicity are taken into account. Also, the first 1 ms or so are omitted as they contain the majority of the effect of the spectral envelope due to $H(f)$.

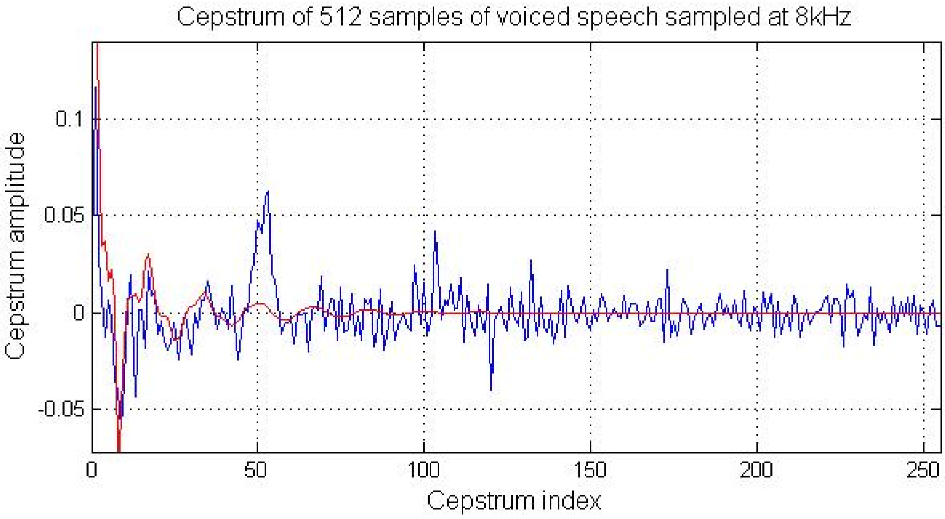
The CPP is now defined as the difference in amplitude between the cepstral peak and the regression line at the same cepstral time (quefrency) [HH96, ARJ$^+$10]. Algorithms have been published for computing CPP, and others, including that by Kay-Pentax [Kay96] have been made available. It is a widely used measure and will be used in this thesis in the form defined by KayPentax within ADSV. The code for this implementation has not been published.

Awan et al [ARJ$^+$10] found strong correlation between cepstral and spectral measure with perceptual severity rating (overall severity). They found CPP, CPP sd, L/H spectral ratio, L/H spectral ratio sd accounted for 90% of the variability for sustained vowels in study by [ARJ$^+$10] while CPP, L/H spectral ratio sd and L/H spectral account for 73% of the variability for connected speech [AR09].

It is interesting to consider how CPP is likely to be affected by other forms of aperiodicity, such as jitter and shimmer. Figure 4.4a shows the same synthetic speech as in Figure 4.1a with the same amount of noise as in Figure 4.3a, but with the fundamental frequency increasing from 143 Hz to 167 Hz. It is clear that there is loss of CPP prominence both due to the additive noise and the spectral spreading arising from the frequency modulation. Clearly a precise measurement of fundamental frequency is now more difficult, and with a little more noise it may become impossible to locate the cepstral peak.



(a) With noise and jitter.



(b) Cepstrum for figure 4.4a

Figure 4.4: Synthetic vowel with noise and jitter Time domain (a) and Cepstrum (b). The time-domain sample index refers to the speech sampled at 8 kHz. The Cepstrum index k represents time in units of 512/ F_s= 64 ms. Therefore k represents (64k) ms in time (or quefrequency). The red graph is the impulse-response

Concerns about the effects of non-stationarity with the CPP method are similar to those discussed earlier with the FFT. These will be borne in mind in future work.

Finally, for completeness, we investigate the effect of amplitude variation on CPP.

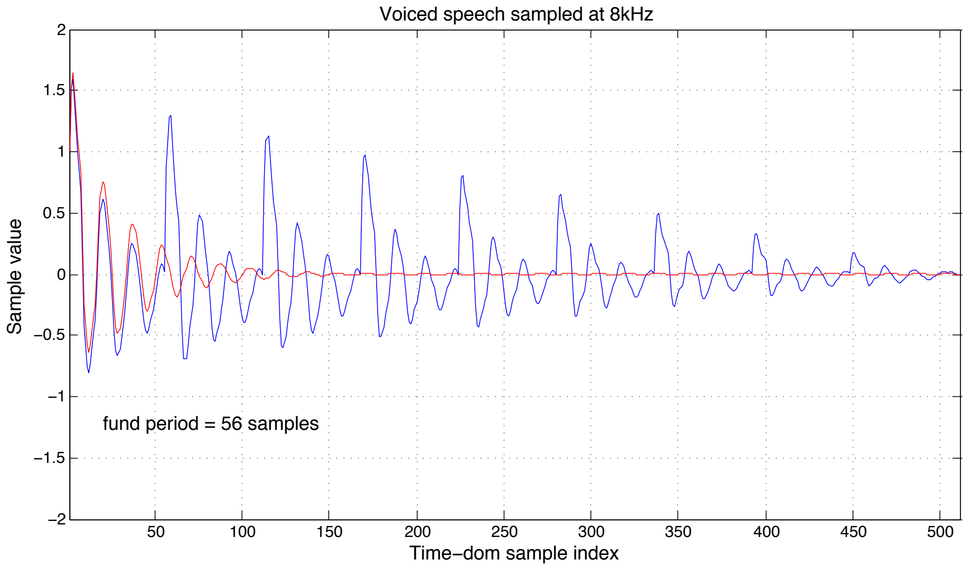


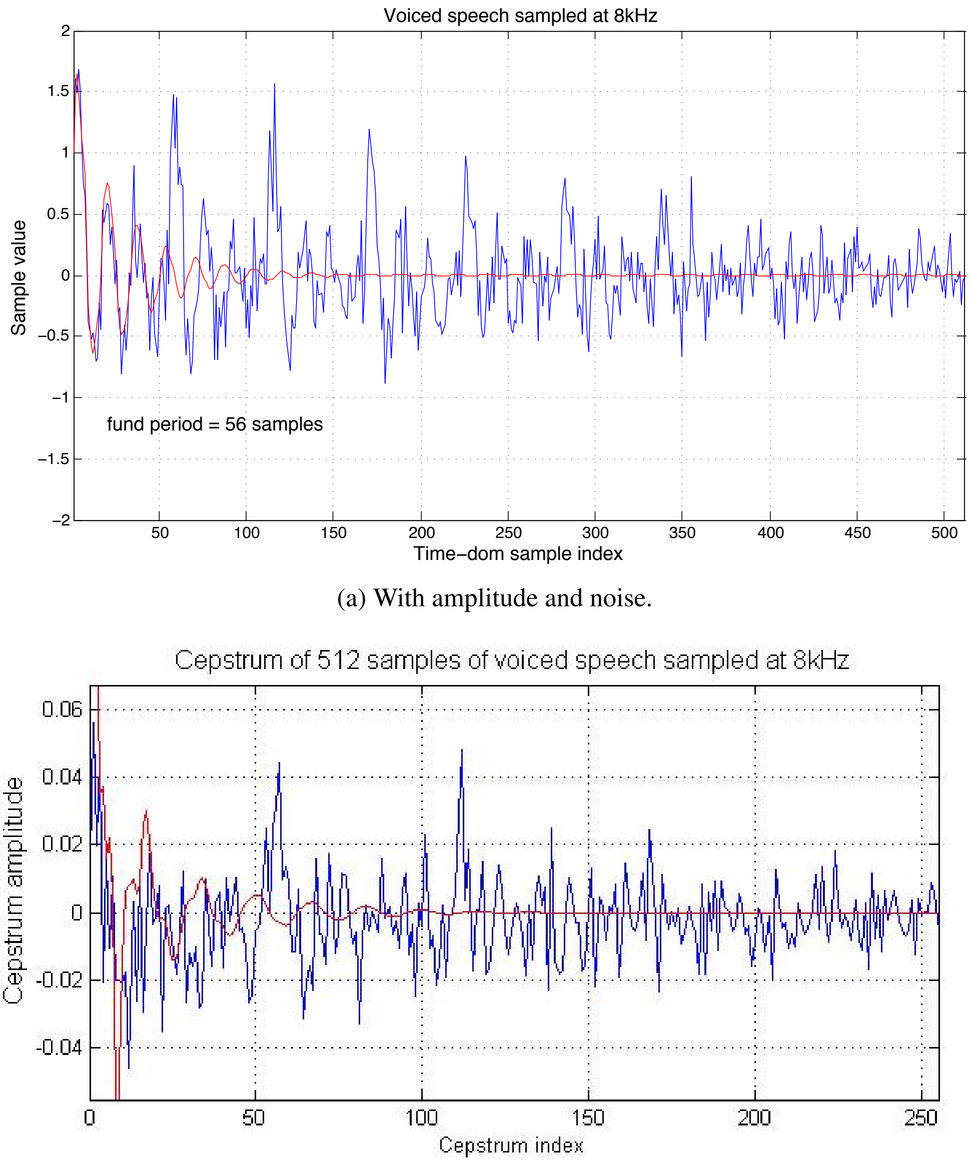
Figure 4.5: Synthetic vowel with amplitude modulation. The time-domain sample index refers to the speech sampled at 8 kHz. The red graph is the impulse-response.

Figure 4.5 shows the original synthetic speech with amplitude modulation applied to the vocal tract excitation. There is no frequency modulation. Figure 4.6a shows the amplitude modulated speech with noise added as in the earlier experiments. The cepstrum obtained is shown in Figure 4.6b. It may be seen that although the noise is more prominent in comparison with the overall speech energy (because the speech amplitude is reducing) the cepstral peak is still quite prominent, the CPP will be quite high and the fundamental frequency can still be accurately determined. These examples illustrate that the effects of frequency modulation (jitter) on measurements of CPP are likely to be more serious than the effects of amplitude modulation.

4.3.4 Autocorrelation based methods

The autocorrelation function of a speech frame $\{s[n]\}_{1,\mathrm{N}}$ is a function of delay d measured in samples. Assuming that the mean value of $\{s[n]\}_{1,\mathrm{N}}$ is zero, it may be estimated as:

$$auto(d) = \frac{N \sum_{n=1}^{N-d} s[n]s[n+d]}{(N-d) \sum_{n=1}^{N} (s[n])^2} \tag{4.4}$$



(a) With amplitude and noise.

(b) Cepstrum for figure 4.6a

Figure 4.6: Synthetic vowel with amplitude and noise time domain (a) and cepstrum (b). The time-domain sample index refers to the speech sampled at 8 kHz. The Cepstrum index k represents time in units of 512/ F_S= 64 ms. Therefore k represents (64k) ms in time (or quefrequency). The red graph is the impulse-response.

It may be shown that -1 $\leqslant auto(d) \leqslant$ 1 and that auto(d) = 1 if s[n] is periodic with period d samples, where d>0. Therefore a valid way of estimating the fundamental frequency of a purely periodic segment of speech is to evaluate auto(d) over a suitable range of values for d and to search for the value of d that makes auto(d) =1. If auto (d) = 1, then it follows that auto (2d)=1 also. Therefore we must take the minimum value of d that makes auto(d)=1. If there is no such value, it may be deduced that $\{s[n]\}_{\mathrm{1,N}}$ is not periodic. However, if we search for the minimum value of d that maximizes auto(d), with $d>0$, and we find that auto(d) is close to 1, we may deduce that $\{s[n]\}_{\mathrm{1,N}}$is approximately periodic and affected by some small degree of aperiodicity due to added noise, frequency or amplitude modulation, or some other reason. If the maximum value of auto(d) is not close to 1, this may indicate that the signal is not close to being periodic and most likely corresponds to unvoiced speech.

The autocorrelation function can therefore be used to detect periodicity or the absence of periodicity (voiced/unvoiced detection). It can also be used to estimate the fundamental frequency of voiced speech. It is a good method, but it has disadvantages largely due to the range over which it is calculated. The fundamental frequency and amplitude of speech cannot be expected to remain exactly the same even over a frame-length of 20 ms or more. The effect of these variations on the shape of the autocorrelation function is difficult to predict and can make voiced/unvoiced decisions and fundamental frequency detection quite difficult. An alternative approach, sometimes referred to as the cross-correlation method, was commonly used in speech coding [KKK90] and is preferred in this thesis for reasons that will be explained.

4.3.5 Cross-correlation methods

Given a speech segment of length N : $\{s[n]\}_{\mathrm{1,N}}$, the basic idea is to extract abutting sub-segments $\{\mathrm{s[n]}\}_{\mathrm{1,L}}$ and $\{\mathrm{s[n]}\}_{\mathrm{L+1,2L}}$ for various values of L. Let $\{\mathrm{x[n]}\}_{\mathrm{1,L}}$ denote $\{\mathrm{s[n]}\}_{\mathrm{1,L}}$ and let $\{\mathrm{y[n]}\}_{\mathrm{1,L}}$ denote $\{\mathrm{s[n]}\}_{\mathrm{L+1,2L}}$. The cross-correlation method searches for the value of L for which $\mathrm{x[n]}_{\mathrm{1,L}}$ and $\{A \times y[n]\}_{\mathrm{1,L}}$ are most similar when A is a scaling factor for the second sub-segment. In one version of this method, the constant A is chosen to maximize the similarity between $\{\mathrm{x[n]}\}_{\mathrm{1,L}}$ and $\{\mathrm{y[n]}\}_{\mathrm{1,L}}$ for any given value of L. A simpler version fixes A to be equal to one.

The reason we introduce A is to try reduce the effect of increasing or decreasing amplitude on our measure of periodicity. The amplitude envelope of voiced speech will be constantly changing especially at the on-set of words and at their ends. Let $e[n] = x[n] - Ay[n]$ for n = 1,2,..., L. Then we must search for the value of L that minimises:

$$E(L) = \frac{1}{L}\sum_{n=1}^{L} e[n]^2 = \frac{1}{L}\sum_{n=1}^{L}(x[n] - Ay[n])^2 \tag{4.5}$$

For any given value of L, we can find the best value of A by differentiating:

$$\frac{dE(L)}{dA} = \frac{1}{L}\sum_{n=1}^{L} -2(x[n] - Ay[n])y[n] \tag{4.6}$$

Setting this to zero to minimise E(L), we get:

$$A = \frac{\sum\limits_{n=1}^{L} x[n]y[n]}{\sum\limits_{n=1}^{L} (y[n])^2} \tag{4.7}$$

It follows that for any value of L:

$$\begin{aligned}
E(L) &= \frac{1}{L}\sum_{n=1}^{L}(x[n])^2 - \frac{2A}{L}\sum_{n=1}^{L}x[n]y[n] + \frac{A^2}{L}\sum_{n=1}^{L}(y[n])^2 \\
&= \frac{1}{L}\sum_{n=1}^{L}(x[n])^2 - \frac{2(\sum\limits_{n=1}^{L} x[n]y[n])^2}{L\sum\limits_{n=1}^{L}(y[n])^2} + \frac{(\sum_{n=1}^{L} x[n]y[n])^2}{L(\sum\limits_{n=1}^{L}(y[n])^2)^2}\sum_{n=1}^{L}(y[n])^2 \\
&= \frac{1}{L}\sum_{n=1}^{L}(x[n])^2 - \frac{(\sum\limits_{n=1}^{L} x[n]y[n])^2}{L\sum\limits_{n=1}^{L}(y[n])^2} \\
&= \frac{1}{L}\sum_{n=1}^{L}(x[n])^2\left[1 - \frac{(\sum\limits_{n=1}^{L} x[n]y[n])^2}{\sum\limits_{n=1}^{L}(x[n])^2\sum\limits_{n=1}^{L}(y[n])^2}\right] = \frac{1}{L}\sum_{n=1}^{L}(x[n])^2(1 - C(L)^2) \\
&\textit{where} \quad C(L) = \frac{\sum\limits_{n=1}^{L}(x[n]y[n])}{\sqrt{\sum\limits_{n=1}^{L}(x[n])^2\sum\limits_{n=1}^{L}(y[n])^2}}
\end{aligned} \tag{4.8}$$

We search for the best value of L to minimise E(L) with positive C(L). If $\{x[n]\}_{1,\mathrm{L}}$ is identical to $\{y[n]\}_{1,\mathrm{L}}$ for some value of L, the signal will be purely periodic, at least over the first 2L samples of the speech frame, and the minimum value of E(L) will be zero. In this case, the maximum value of C(L) over all L, call this $C_{\max}$ will be equal to 1. If $\{x[n]\}_{1,\mathrm{L}}$ is identtical to $\{-y[n]\}_{1,\mathrm{L}}$ for some value of L, the signal is not necessarily purely periodic, though the minmum value of E(L) will be zero with

C(L) equal to -1. In speech processing, such negative correlation is disregarded as it does not indicate periodicity. If a value of L cannot be found such that C(L) is exactly equal to 1, strictly speaking s[n] is not periodic. In strict terms, a signal is either periodic or not periodic. But if the maximum positive value of C(L) is close to 1, it may be argued that there is 'a degree of periodicity' in $\{s[n]\}$. Trying to quantify this 'degree of periodicity' is quite difficult. We have found it useful and justifiable to define the maximum obtainable value of C(L) as the 'degree of periodicity' and the corresponding value of L the 'period'. This extends the strict normal definition of 'period'. We avoid terms like 'pseudo-period' and 'pseudo-periodicity' since they have come to be associated with concepts of short-term stationarity which are too specific.

4.3.6 Measuring degrees of periodicity using the cross-correlation

If $C_{\max}$is defined as the degree of periodicity, we can define $(1 - C_{\max}$) as the degree of non-periodicity, or 'aperiodicity index'. Speech is rarely exactly periodic even when amplitude variations are disregarded. But it can be very close to periodic during voiced speech and highly aperiodic during unvoiced speech. Assuming unvoiced speech to be sourced by a spectrally white turbulent excitation (often loosely termed 'white noise') the maximum value of C(L) can be quite low, typically around 0.2. One may expect $C_{\max}$ to approach zero, but the finiteness of the sample means that we cannot expect to obtain zero exactly. However, even strongly aperiodic consonants can be spectrally 'coloured' to some extent by vocal tract resonance and hence have some small degree of periodicity. Strongly periodic voiced sounds may have an element of turbulent flow, or breathiness, causing a small degree of aperiodicity. Such aperidiocity may have different causes all of which are of great interest to speech analysis. It may be caused by turbulent flow when the vocal cords do not close completely within each pitch-cycle. Or it may be caused by frequency modulation (Jitter or vibrato) or amplitude (shimmer).

Aperiodicity within voiced speech may be considered to be caused by the addition of zero mean white noise $\{N[n]\}$ of variance σ^2. In some cases this may be the true cause, but in other cases it may be a convenient assumption for modeling the true situation. In either case, we can find the value of L which maximises C(L) and then express:

$$\begin{aligned} x[n] &= p[n] + N_{\mathrm{x}}[n] \\ y[n] &= p[n] + N_{\mathrm{y}}[n] \end{aligned} \tag{4.9}$$

for n=1,2,....,L where p[n] is one cycle of some periodic signal of period L samples, and $N_x[n]$ and $N_y[n]$ are zero mean white noise signals, extracted from {N[n]} and therefore of equal power with zero correlation between them. Thus, C_{max} now becomes:

$$\begin{aligned} C_{\max} &= \frac{\frac{1}{L}\sum_{n=1}^{L}(p[n]+N_{\mathrm{x}}[n])(p[n]+N_{\mathrm{y}}[n])}{\sqrt{\frac{1}{L}\sum_{n=1}^{L}(p[n]+N_{\mathrm{x}}[n])^2\frac{1}{L}\sum_{n=1}^{L}(p[n]+N_{\mathrm{y}}[n])^2}} \\ &\approx \frac{\frac{1}{L}\sum_{n=1}^{L}(p[n])^2}{\sqrt{\frac{1}{L}\sum_{n=1}^{L}(p[n]^2+N_{\mathrm{x}}[n]^2)\frac{1}{L}\sum_{n=1}^{L}(p[n]^2+N_{\mathrm{y}}[n]^2)}} \end{aligned}$$

since N $_x$[n] and N $_y$[n] are uncorrelated with each other and with p[n]. Therefore

$$\begin{aligned} C_{\max} &\approx \frac{\frac{1}{L}\sum_{n=1}^{L}(p[n])^2}{\sqrt{(\frac{1}{L}\sum_{n=1}^{L}(p[n])^2)^2+2\frac{1}{L}\sum_{n=1}^{L}(N_{\mathrm{x}}[n])^2\frac{1}{L}\sum_{n=1}^{L}(p[n])^2+\frac{1}{L}\sum_{n=1}^{L}(N_{\mathrm{x}}[n])^2\frac{1}{L}\sum_{n=1}^{L}(N_{\mathrm{y}}[n])^2}} \\ &= \frac{(\sum_{n=1}^{L}p[n])^2}{\sqrt{\sum_{n=1}^{L}(p[n]^2)^2+2\sum_{n=1}^{L}N_{\mathrm{x}}[n]^2\sum_{n=1}^{L}(p[n])^2+(\sum_{n=1}^{L}N_{\mathrm{x}}[n]^2)^2}} \\ &= \frac{1}{\sqrt{1+2\sum_{n=1}^{L}N_{\mathrm{x}}[n]^2/\sum_{n=1}^{L}p[n]^2+(\sum_{n=1}^{L}N_{\mathrm{x}}[n]^2/\sum_{n=1}^{L}p[n]^2)^2}} \\ &= \frac{1}{\sqrt{(1+\sum_{n=1}^{L}(N_{\mathrm{x}}[n])^2/\sum_{n=1}^{L}(p[n])^2)^2}} = \frac{1}{1+\sum_{n=1}^{L}(N_{\mathrm{x}}[n])^2/\sum_{n=1}^{L}(p[n])^2} = \frac{1}{1+1/HNR} \end{aligned} \tag{4.10}$$

where HNR is the ‘harmonic-to-noise ratio’ defined as:

$$HNR = \sum_{n=1}^{L} (p[n])^2 / \sum_{n=1}^{L} (N_x[n])^2 \tag{4.11}$$

Therefore,

$$\begin{aligned} 1/HNR &\approx 1/C_{\max} - 1 \\ &\approx (1 - Cmax)/Cmax \end{aligned} \tag{4.12}$$

which means that

$$HNR \approx Cmax/(1 - Cmax) \tag{4.13}$$

This estimation formula for HNR has been tested for additive white noise by means of a MATLAB simulation program. This program adds zero mean uniformly distributed white noise to a periodic signal of fundamental frequency 200 Hz sampled at 40 kHz. The period is therefore 200 samples. The program was run for a fixed periodic signal power, and increasing levels of additive noise giving a signal to noise ratio ranging from about 6 dB to 30 dB. It may be seen from the graph below that the supposed HNR formula predicts the true Signal to Noise Ratio (SNR) level quite accurately when the aperiodicity is really due to additive white noise. The maximum error that occurred was less than 1 dB and the variance of the difference between predicted and true value of SNR was 0.004.

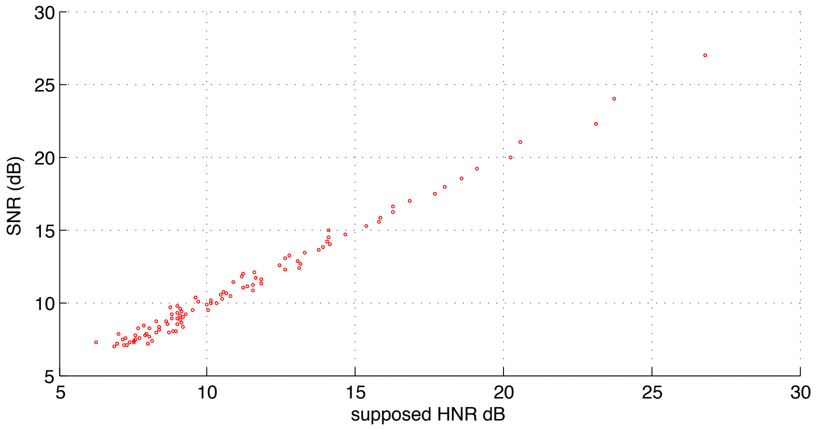


Figure 4.7: HNR measurements for a periodic signal of fundamental frequency 200 Hz sampled at 40 kHz waveform with added noise compared with SNR

The cross-correlation technique as used above is different from the more conventional ‘autocorrelation’ technique. It looks for the cross-correlation between consecutive pitch-cycles rather than peaks in an autocorrelation function calculated across a fixed duration speech frame containing many cycles. It behaves better than the autocorrelation method when the frame is in transition, i.e. when characteristics are changing rapidly. In optimising the value of the scaling factor A, it tries to cancel out the effect of amplitude changes which include shimmer and also the changing envelope at the onset or endings of phonemes. Optimising A has advantages for estimating HNR, jitter and voicing decisions. However, it was discovered that difficulties are created by optimising A when the aim is to estimate the fundamental frequency. A difficulty that can arise is the mistaking of short term periodicity due to vocal tract resonances (formants) for the longer term pitch-cycle periodicity due to vocal cord vibration. The short term periodicity creates peaks in the cross-correlation function which are enhanced by the optimisation of A. Essentially, the optimisation of A can cancel out the decay in amplitude of a resonance due to a formant (usually the first formant) and can thus make a decaying sinusoid look like a constant sinusoid. The constant sinusoid then gives a higher measure of cross-correlation than is appropriate.

Fortunately, the solution to this problem is straightforward. For fundamental frequency detection we use the simpler ‘constant A’ version of the cross-correlation method for fundamental frequency detection, while retaining the use of the ‘optimised A’ version for all other measurements. It is easily shown, by manipulating equation (4.8), that fixing A to be equal to one gives the following formula for mean-squared error E(L) and cross-correlation value C(L):

$$\begin{aligned}E(L) &= \frac{1}{L}\sum_{n=1}^{L}(x[n])^2 - \frac{2}{L}\sum_{n=1}^{L}x[n]y[n] + \frac{1}{L}\sum_{n=1}^{L}(y[n])^2 \\ &= \frac{1}{L}\sum_{n=1}^{L}((x[n])^2+(y[n])^2)\left(1-\frac{\sum\limits_{n=1}^{L}x[n]y[n]}{(\sum\limits_{n=1}^{L}(x[n])^2+(y[n])^2)/2}\right) \\ &= \frac{1}{L}\sum_{n=1}^{L}(((x[n])^2+(y[n])^2)(1-C(L)) where \quad C(L) = \frac{\sum\limits_{n=1}^{L}x[n]y[n]}{\left(\sum\limits_{n=1}^{L}(x[n])^2+(y[n])^2\right)/2}\end{aligned} \tag{4.14}$$

It was found that this simplification to the cross-correlation technique as originally defined greatly reduces the occurrences of fundamental frequency estimation errors for the reason explained above. It is only used for fundamental frequency estimation.

The accuracy of the cross-correlation technique for estimating the fundamental frequency of voiced speech waveforms was verified by applying to it synthesised vowel sounds with known fundamental frequency and simulated formant characteristic of known vowel sounds. Table D.1 in appendix D shows estimate of fundamental frequency for synthesised vowels, i.e. /a/, /o/ and /e/ over the range 120 Hz to 200 Hz as estimated by the thesis software. The thesis measurements for three synthesised vowels is close to the nominal values of fundamental frequency. The maximum discrepancy is between the fundamental frequency 200 Hz with vowel /o/ with frequency 199.53 Hz that is about 0.23%.

4.3.7 Aperiodicity Index (API) and harmonic to noise ratio (HNR)

The purpose of this chapter is to consider how best to measure and quantify features which characterise speech in the five GRBAS dimensions. The degree to which a speech segment is periodic or aperiodic is a clear predictor of 'grade' (G) and 'roughness' (R), and the harmonic-to-noise ratio (HNR) as defined in the previous section is clearly relevant to 'breathiness' (B). The Aperiodicity Index (API) is defined as 1 - $\mathrm{C_{max}}$ as calculated for the value of L that maximizes C as defined above. The corresponding value of L is referred to as the 'period' even though the speech segment may not be purely periodic. The HNR is intended to indicate the degree to which a purely periodic waveform may have been affected by additive white noise. If the signal really is a periodic signal affected by additive white noise, HNR gives a reliable estimate of the signal-to noise ratio. Otherwise, HNR may be considered a model of the true situation. We have shown that a reliable estimate of HNR is given by the Equation (4.13).

4.3.8 Enhanced cross-correlation method

An improvement of the cross-correlation function replaces the constant A which multiplies the second abutting segment {y[n]} by the time varying function An+B. Instead of choosing just A, we now try to choose both A and B to maximise the similarity between $\{x[n]\}_{0,\mathrm{L}}$ and $\{(A+nB)y[n]\}_{0,\mathrm{L}}$. Clearly this allows {y[n]} to be scaled up or down by a sequence of values that decrease linearly with time at the onset of vowels and increase linearly with time as the envelope decays at the ends of vowels. Defining:

$$E(L) = \frac{1}{L}\sum_{n=1}^{L} e[n]^2 = \frac{1}{L}\sum_{n=1}^{L}(x[n] - (A+nB)y[n])^2 \tag{4.15}$$

for any given value of L, we can find the best value of A and B by differentiating to obtain:

$$\frac{dE(L)}{dA} = \frac{1}{L}\sum_{n=1}^{L} -2(x[n] - (A+nB)y[n])y[n] \tag{4.16}$$

$$\frac{dE(L)}{dB} = \frac{1}{L}\sum_{n=1}^{L} -2n(x[n] - (A+nB)y[n])y[n] \tag{4.17}$$

Setting both these expressions to zero to minimise E(L), we get the matrix equation:

$$\begin{bmatrix} \sum_{n=1}^{L}(y[n])^2 & \sum_{n=1}^{L} n(y[n])^2 \\ \sum_{n=1}^{L} n(y[n])^2 & \sum_{n=1}^{L} n^2(y[n])^2 \end{bmatrix} \times \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{L} x[n]y[n] \\ \sum_{n=1}^{L} nx[n]y[n] \end{bmatrix}$$

This second order matrix equation is easily set up and solved, for example in MATLAB, to obtain the best values of A and B for any given L. A maximum value of C can then be obtained and used as above. This value of $\mathrm{C_{max}}$ will be even closer to 1 at onsets and endings of vowels if the further reduction of the effect of amplitude modulation has been successful.

4.3.9 Advantages of the cross-correlation method

As defined above, the cross-correlation method is really a modified form of the autocorrelation method since the two segments compared are extracted from the same speech segment. However the modification is beneficial in giving an instantaneous measurement of periodicity that is unaffected by amplitude and fundamental frequency variation that occur within the complete frame as analysed by the autocorrelation method. Hence it may be expected to give a good estimate of the effect of added noise on periodicity. The effects of frequency and amplitude modulation are better estimated separately using standard definitions of jitter, vibrato and shimmer.

4.4 Measuring fundamental frequency, voicing, jitter and shimmer

The cross-correlation technique relies on the correlation between successive waveform segments to determine the most likely value of fundamental frequency (F_0). This is essentially ‘waveform matching’ as discussed in Chapter 2. The wave-shapes of successive pitch-cycle candidates must be maximally similar, i.e. the mean square difference between them must to minimised, for the candidates to be taken seriously. There are many finer points to be considered before a definite decision about F_0 can be taken. This is because shorter term periodicity due to vocal tract resonance may be mistaken for F_0, and also longer term periodicity at sub-multiples of F_0, especially half and one third of F_0, will always exist when there is periodicity at F_0. It is quite common for a cross-correlation peak at $0.5 \times F_0$ to be higher than that at F_0, especially with speech affected by additive random components. The logic of deciding which cross-correlation peak belongs to F_0 can be quite complicated.

Detecting F_0 is a necessary precursor to calculating many other speech parameters, including jitter, shimmer and harmonic to noise ratio. The same process allows the voiced/unvoiced decision required for analysing connected speech. Pitch doubling is a common error, though fortunately, as will be argued later, the effect of this on measurements of relative shimmer and jitter, and also harmonic-to-noise ratio, may not be too serious.

Jitter and shimmer were introduced as features of voiced speech in Chapter 2. Both features have been widely used in clinical and scientific settings for detecting voice pathologies, and there are strong reasons for believing that they may help to characterize speech in the five GRBAS dimensions [Wag13, KG05]. Breathy and rough voices will have measurable degrees of jitter and shimmer during voiced speech. Jitter and shimmer are often measured for sustained vowels, though the analysis of voiced parts of connected speech is possible with a reliable voiced/unvoiced detection mechanism. According to MDVP [MDV], values of jitter and shimmer above certain thresholds, which MDVP specifies, are considered to be indicative of pathological voices. However, even normal voices exhibit cycle-to-cycle pitch and amplitude perturbations [Dav79, IVL70]

Different formulae have been proposed for measuring jitter and shimmer as discussed in Chapter 2. Five well known formulae are presented for jitter (2.1) to (2.5), four for shimmer (2.6) to (2.9) and also there is a range of user-adaptable versions in MDVP, suitable for sustained vowels (2.10) to (2.12). Jitter and shimmer must be distinguished from the frequency and amplitude modulation that is due to natural intonation and this consideration has led to the range of different formulae. When measuring jitter and shimmer care has to be taken that the resolution in amplitude and frequency is sufficiently fine, and this can require the up-sampling of the waveform. In this work, all recordings use a sampling rate of 44.1 kHz, with 16 bits/sample uniform quantisation. We believe that this digitisation process offers sufficient accuracy without up-sampling.

The formulae for jitter and shimmer require cycle-to-cycle measurements of F_0. Perturbation features may be strongly affected by the difficulty of determining a pitch-frequency in significantly dysphonic voices. As discussed in Chapter 2, differences between the Praat and MDVP software suites in the way F_0 are derived is responsible for significant differences in the values of jitter and shimmer obtained, even when the essential formulae are identical. After studying the differences it was concluded that the ‘waveform matching’ approach [Boe09] used by Praat is likely to be the most reliable. This has been adopted by the software produced in this thesis.

With severe damage, there may indeed be little or no periodicity with the voice becoming hoarse, or whispered, due to the excitation being entirely produced by turbulent air-flow. Such cases must be catered in the DSP analysis. We cannot allow the feature detection process for voiced speech to simply fail when periodicity cannot be detected. Where the analysis is done both on sustained vowels and connected speech, the latter is likely to be both more difficult to process and less discriminating when comparing normal and pathological voice [ZJ08].

4.5 Measuring jitter for artificial voiced speech

To evaluate the ‘thesis software’ for jitter estimation and compare it with the well known Praat and MDVP software suites, samples of artificial voiced speech were produced by exciting an all-pole vocal tract model, with glottal pulse shaping and lip-radiation filtering, by a periodic series of discrete time impulses. The poles were given radii of 0.992, 0.99, 0.988, and 0.986 with frequencies ±610, ±1300, ±2450 and ±3600 Hz respectively to emulate the phoneme /a/. The sampling rate was 44.1 kHz. The required pole radii are about 5.5 times closer to 1 than would be the case with an 8 kHz sampling rate. Jitter was synthesised by introducing Pitch- Period Variation (PPV) into the time locations of the excitation impulses. There was no simulated shimmer or

PPV%	th-RL%	th-RAP%	th-PP5%	Pra-RL%	Pra-RAP%	Pra-PPQ5
0.0	0.00	0.00	0.00	0.00	0.00	0.00
0.2	0.20	0.13	0.16	0.19	0.11	0.14
0.4	0.50	0.30	0.34	0.45	0.26	0.30
0.6	0.63	0.37	0.38	0.55	0.32	0.33
0.8	1.03	0.59	0.68	0.90	0.52	0.59
1.0	1.13	0.65	0.68	1.00	0.57	0.56
1.2	1.48	0.94	0.91	1.31	0.82	0.81
1.4	1.59	0.98	0.92	1.44	0.87	0.83
1.6	1.71	1.07	1.00	1.50	0.93	0.93
1.8	2.11	1.25	1.26	1.92	1.15	1.16
2.0	1.94	1.15	1.27	1.80	1.08	1.25
2.2	2.06	1.19	1.45	1.92	1.12	1.33
2.4	2.83	1.61	1.98	2.71	1.57	1.93
2.6	2.94	1.75	1.79	2.89	1.55	1.55
2.8	3.23	1.77	2.12	3.09	1.66	2.04
3.0	2.80	1.62	1.69	2.65	1.50	1.59
3.2	3.11	1.87	1.85	2.86	1.73	1.79
3.4	3.84	2.18	2.59	3.73	2.14	2.56
3.6	4.20	2.51	2.92	3.98	2.36	2.83
3.8	4.82	2.92	3.10	4.51	2.66	2.95
4.0	4.70	2.70	3.11	4.52	2.60	3.00
4.2	4.48	2.63	3.29	4.34	2.47	3.09
4.4	4.71	2.59	3.49	3.81	2.10	2.97
4.6	5.41	3.20	3.80	3.85	2.07	2.91
4.8	4.53	2.82	2.98	3.44	2.21	2.61
5.0	4.96	2.85	3.41	4.07	2.29	2.93
5.2	5.04	2.81	3.69	3.98	2.13	3.33
5.4	5.45	3.13	3.73	5.32	2.97	3.57
5.6	6.17	3.65	3.68	4.94	3.06	3.07
5.8	6.49	3.84	4.19	4.29	2.52	3.23
6.0	7.84	4.82	4.86	4.97	3.30	4.25
6.2	6.16	3.60	3.74	5.60	3.27	3.69
6.4	7.18	4.44	4.61	5.62	3.39	4.41
6.6	7.53	4.37	4.70	4.48	2.47	3.45
6.8	8.25	4.75	5.76	6.56	3.78	5.05
7.0	9.04	5.43	5.38	6.01	2.59	3.08
7.2	8.20	5.18	4.58	5.32	3.07	3.02
7.4	8.72	4.98	5.99	6.30	3.50	4.52
7.6	7.26	4.09	5.26	5.81	3.42	4.15
7.8	7.83	4.64	5.37	5.79	3.27	4.12
8.0	9.34	5.56	6.01	5.83	3.84	6.32

Table 4.1: Comparison of RL Jitter measurements for artificial voiced speech

added noise in this experiment. Table 4.1 shows three commonly used estimates of jitter, i.e. Relative-Local (RL), RAP and PPQ5 as estimated by the thesis software and also Praat. These estimates are defined by Equations 2.2 to 2.4. The results for RL-jitter as estimated by thesis and Praat software are presented graphically in Figure 4.8. The 'thesis' and Praat measurements are both close to each other and to the nominal values of RL jitter up to about 4%. For nominal values of jitter greater than about 4% the 'thesis' and Praat estimates of RL jitter diverge, though the thesis estimates remain closer to the nominal values than the Praat estimates. Similar comparisons are observed for the other estimates of jitter (RAP and PPQ5). MDVP was unable to give reasonable estimates of jitter for these artificial speech files.

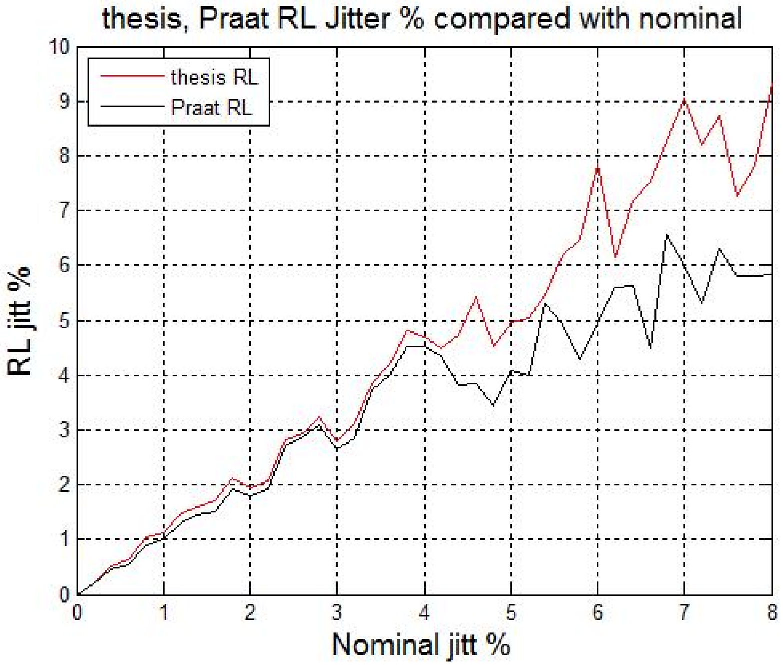


Figure 4.8: Comparison of RL Jitter measurements for artificial voiced speech for thesis software and Praat software

4.6 Measuring shimmer for artificial voiced speech

To evaluate the 'thesis software' for shimmer estimation and to compare it with the Praat and MDVP software suites, samples of artificial voiced speech were produced by exciting the same all-pole vocal tract model, with glottal pulse shaping and lip-radiation filtering, as was used in the previous section. Shimmer was synthesised by introducing variation (Shimmer Variation (SHV)) into the amplitudes of the excitation impulses. There was no simulated jitter or added noise in this experiment. Table 4.2 shows three commonly used estimates of shimmer, i.e. relative-local (RL), APQ3 and APQ5 as estimated by the thesis software and also by Praat.

These estimates are defined by Equations (2.6), (2.9) and (2.8). The results for RL-shimmer (Equation 2.6) as estimated by 'thesis' and Praat software are presented graphically in Figure 4.9. The 'thesis' and Praat measurements are both close to each other and to the nominal values of RL shimmer. Similar comparisons are observed for the other estimates of shimmer (APQ3 and APQ5). MDVP was unable to give reasonable estimates of shimmer for these artificial speech files.

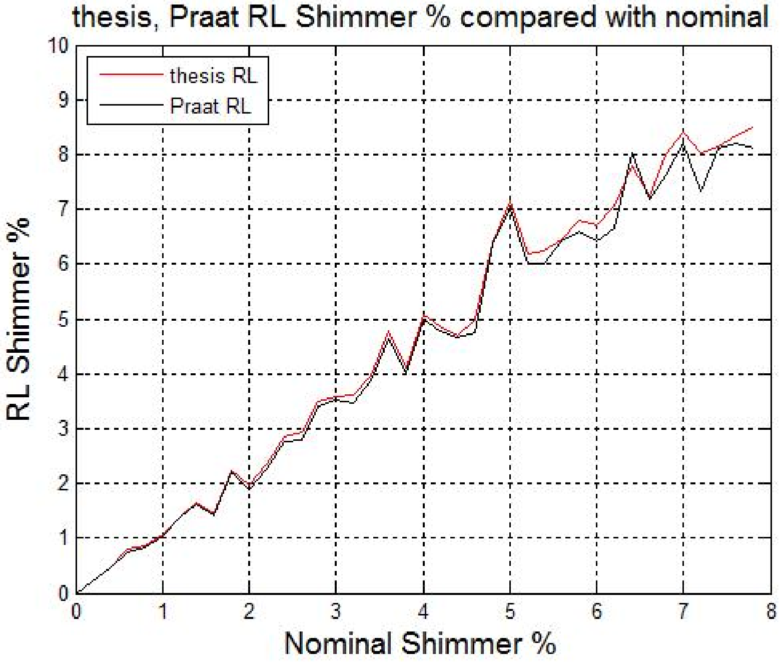


Figure 4.9: Comparison of RL Shimmer measurements for artificial voiced speech for thesis software and Praat software

4.7 Measuring artificial voiced speech with jitter and shimmer

So far the 'thesis software' has been evaluated and compared with the Praat software suites for jitter and shimmer when they occur independently, and when there is no noise due to turbulent air-flow. In this section we investigate to what extent jitter, shimmer and HNR can be measured when they occur simultaneously. Samples of

SHV%	th-RL%	th-APQ3%	th-APQ5%	Pra-RL%	Pra-APQ3%	Pra-APQ5%
0.0	0.00	0.00	0.00	0.00	0.00	0.00
0.2	0.22	0.13	0.16	0.22	0.12	0.15
0.4	0.45	0.26	0.29	0.44	0.25	0.28
0.6	0.80	0.47	0.53	0.76	0.44	0.52
0.8	0.87	0.53	0.54	0.84	0.51	0.53
1.0	1.06	0.59	0.71	1.03	0.58	0.69
1.2	1.40	0.84	0.88	1.38	0.82	0.87
1.4	1.65	0.96	1.06	1.61	0.92	1.05
1.6	1.46	0.77	0.95	1.42	0.75	0.93
1.8	2.23	1.30	1.42	2.20	1.26	1.37
2.0	1.97	1.11	1.19	1.88	1.05	1.17
2.2	2.34	1.27	1.38	2.26	1.24	1.36
2.4	2.86	1.66	1.77	2.76	1.59	1.74
2.6	2.93	1.62	1.86	2.80	1.55	1.83
2.8	3.48	2.01	2.15	3.41	1.98	2.11
3.0	3.58	2.12	2.19	3.52	2.05	2.17
3.2	3.61	2.12	2.25	3.47	2.03	2.19
3.4	3.95	2.28	2.75	3.88	2.24	2.69
3.6	4.78	2.79	3.06	4.63	2.70	3.03
3.8	4.11	2.37	2.58	4.01	2.29	2.51
4.0	5.06	2.99	3.58	4.99	2.94	3.55
4.2	4.87	2.89	3.00	4.79	2.82	2.88
4.4	4.70	2.83	3.08	4.65	2.76	3.01
4.6	5.00	2.78	3.35	4.76	2.63	3.28
4.8	6.40	3.71	4.39	6.36	3.62	4.26
5.0	7.14	4.31	4.28	7.01	4.24	4.15
5.2	6.19	3.68	3.97	6.00	3.58	3.92
5.4	6.25	3.63	4.79	6.01	3.48	4.73
5.6	6.45	3.64	3.99	6.41	3.53	3.87
5.8	6.81	4.03	4.43	6.59	3.86	4.35
6.0	6.70	3.95	4.54	6.41	3.81	4.48
6.2	7.05	4.12	4.74	6.66	3.81	4.42
6.4	7.78	4.70	5.24	8.02	4.74	5.13
6.6	7.23	4.22	4.44	7.17	4.20	4.25
6.8	8.00	4.80	5.55	7.62	4.57	5.53
7.0	8.41	4.84	5.00	8.21	4.62	4.89
7.2	8.04	4.33	5.41	7.33	3.78	5.33
7.4	8.14	4.65	4.97	8.10	4.64	4.98
7.6	8.36	4.62	5.75	8.20	4.60	5.85
7.8	8.49	5.01	5.35	8.11	4.79	5.41

Table 4.2: Comparison of Shimmer measurements for artificial voiced speech

artificial voiced speech were produced by introducing both jitter (PPV) and shimmer (SHV) into the frequency and amplitudes of the excitation impulses. Firstly no added noise was introduced to produce the results in Table 4.3 for RL jitter and RL shimmer only. Secondly the whole experiment was repeated with additive noise to achieve a nominal signal to noise ratio of 10 dB.

It may be observed that estimates of jitter are largely independent of shimmer and HNR. Similarly, thesis-HNR is largely independent of jitter and only slightly affected by shimmer. The 'thesis' measurements of shimmer, on the other hand, are strongly affected by both Jitter and HNR.

The effect of jitter on shimmer is easily explained. It is due to the interaction between consecutive pitch-periods when the resonance due to one excitation pulse has not died away before the next excitation pulse arrives. The continued oscillation will be added to the next excitation pulse. Without jitter, the added component will tend to be the same for all excitation pulses. But when there is jitter, it will change as the time location of the excitation pulse changes with respect to the previous excitation. Despite much effort, this dependency of shimmer on jitter has not been eliminated in the thesis software as may be observed in Table 4.3. To see the effect clearly, observe the values of 'thesis' shimmer obtained when the nominal shimmer is zero and the nominal jitter increases from 0 to 6%. It may be observed also that the same effect occurs with the Praat estimate of RL shimmer. The dependencies of 'thesis' and 'Praat' shimmer estimates on jitter and HNR are clearly non-linear and are unlikely to be successfully eliminated by Principal Components Analysis (PCA) or more sophisticated versions of PCA. Reducing this dependency would be a useful topic for further research.

Finally, it may be observed in Tables 4.3 and 4.4 that whereas the 'thesis' measurements of HNR remain largely independent of synthesized jitter and shimmer, the Praat measurement of HNR is highly dependent on the levels of both jitter and shimmer. The Praat measurements of HNR reduce remarkably from the known value as levels of jitter and shimmer increase. This is a big surprise and a strong reason for preferring the thesis software despite the constant 1 dB bias in HNR that is discussed in the next section.

4.8 Measuring Harmonic-to-Noise Ratio (HNR)

Many papers have been published on the derivation of HNR by different methods [YGB82, Boe93, AF94, DW03, SBD05, FGHD$^+$09]. The authors of 'Praat' [Pa07]

PPV	SHV	th-jit%	th-shim%	th-HNR	Pra-jit%	Pra-shim %	Pra-HNR
0	0	0.00	0.00	29.98	0.00	0.00	31.82
0	1	0.00	1.09	28.91	0.00	1.05	31.44
0	2	0.00	2.27	28.25	0.00	2.17	29.43
0	3	0.00	2.81	25.49	0.00	2.75	28.51
0	4	0.00	4.07	26.82	0.01	3.91	25.38
0	5	0.00	5.86	24.09	0.01	5.72	24.50
0	6	0.00	6.37	23.84	0.01	6.28	23.38
1	0	1.21	0.51	35.72	1.02	0.90	17.94
1	1	1.35	1.23	39.23	1.19	1.42	16.08
1	2	1.22	2.31	31.31	1.09	2.36	16.74
1	3	1.23	2.98	35.69	1.06	3.00	17.04
1	4	1.07	5.32	26.86	0.94	5.25	17.87
1	5	1.02	6.04	30.98	0.94	6.05	17.25
1	6	1.13	6.89	23.45	0.97	6.92	17.51
2	0	2.21	1.01	31.73	2.06	1.62	11.42
2	1	2.61	1.42	29.68	1.86	2.02	11.63
2	2	2.31	2.45	35.22	2.23	2.74	10.17
2	3	2.15	3.77	33.77	2.02	3.95	11.94
2	4	2.33	3.66	29.26	2.23	3.54	10.99
2	5	2.13	6.02	31.26	1.98	6.10	11.37
2	6	2.31	6.77	36.91	2.09	6.68	11.18
3	0	3.73	1.85	26.88	3.46	2.67	7.69
3	1	3.14	1.99	27.80	2.79	2.80	8.10
3	2	3.17	3.06	29.74	3.04	3.71	7.49
3	3	3.23	3.82	28.26	3.10	4.45	8.46
3	4	3.03	4.97	34.00	2.89	5.25	8.32
3	5	3.40	6.66	27.62	3.21	6.96	7.62
3	6	3.96	7.05	35.77	3.89	7.15	5.97
4	0	3.93	2.76	26.21	3.87	3.74	5.75
4	1	4.35	3.12	28.69	4.20	4.13	5.03
4	2	4.92	3.62	27.72	4.74	4.56	4.27
4	3	5.21	4.53	27.21	4.18	4.98	4.09
4	4	4.52	5.19	27.98	4.18	5.86	4.91
4	5	4.80	6.70	27.17	4.49	7.53	4.33
4	6	4.51	6.27	27.36	4.37	6.45	4.76
5	0	6.57	4.45	26.24	4.69	4.31	4.01
5	1	6.23	4.35	27.71	4.55	6.44	4.80
5	2	5.07	4.11	30.90	3.64	4.60	5.69
5	3	4.98	4.80	44.32	4.87	5.62	3.96
5	4	5.05	5.44	35.73	4.66	6.25	4.58
5	5	5.72	6.95	37.24	5.12	7.91	3.04
5	6	5.12	7.17	34.36	3.96	7.20	5.21
6	0	6.79	4.52	29.27	4.70	4.44	4.71
6	1	6.09	4.58	26.94	5.74	5.42	2.76
6	2	7.22	4.93	30.37	4.11	7.06	5.12
6	3	5.50	5.92	31.68	4.72	6.75	4.49
6	4	6.37	6.53	26.18	4.09	7.14	4.29
6	5	6.98	6.22	26.98	5.02	5.62	3.81
6	6	7.11	8.52	30.52	5.18	7.86	3.16

Table 4.3: Jitter & Shimmer for artificial voiced speech with SNR = infinity

PPV	SHV	th-RLjit%	th-RLshim%	th-HNR dB	Pra-RLjitt%	Pra-RLshim%	Pra-HNR dB
0	0	0.27	5.98	11.24	0.22	2.70	10.26
0	1	0.30	5.57	11.29	0.23	2.59	10.23
0	2	0.21	6.17	11.19	0.23	4.26	10.11
0	3	0.26	7.16	11.30	0.28	4.95	10.16
0	4	0.28	7.62	11.29	0.23	4.96	10.15
0	5	0.28	6.22	11.23	0.24	5.17	10.11
0	6	0.25	7.44	11.40	0.26	6.75	10.09
1	0	0.94	4.77	11.21	0.88	2.32	9.44
1	1	1.09	6.09	11.25	1.03	2.47	9.24
1	2	1.11	6.74	11.29	0.99	2.96	9.44
1	3	1.28	6.52	11.41	1.18	4.02	9.03
1	4	1.03	6.55	11.28	0.97	4.98	9.19
1	5	1.33	8.00	11.28	1.27	5.60	8.73
1	6	1.08	9.83	11.34	1.03	7.17	9.15
2	0	2.18	6.10	11.19	2.05	3.96	7.17
2	1	2.19	5.67	11.02	2.01	2.91	7.21
2	2	2.37	6.42	11.18	2.30	3.20	6.67
2	3	2.21	5.92	11.29	2.17	4.29	7.33
2	4	1.76	6.86	11.18	1.57	5.73	7.86
2	5	2.18	6.86	11.16	2.05	5.84	6.98
2	6	2.20	8.92	11.34	1.99	6.89	7.37
3	0	3.25	6.08	11.04	3.27	4.32	4.93
3	1	3.03	6.27	11.01	2.87	3.79	5.71
3	2	3.31	5.53	11.05	3.16	3.85	4.95
3	3	2.56	8.27	11.07	2.52	5.17	6.16
3	4	3.98	6.05	10.95	3.79	5.15	4.48
3	5	3.98	8.45	11.13	3.69	7.75	4.49
3	6	2.83	7.71	11.28	2.81	6.80	5.60
4	0	4.27	7.12	10.96	3.82	3.85	3.89
4	1	3.81	8.20	11.11	3.79	5.50	4.13
4	2	4.65	7.85	11.08	4.37	5.46	3.04
4	3	3.92	6.66	10.97	3.50	5.05	4.04
4	4	4.38	6.66	10.93	3.91	7.43	4.03
4	5	4.89	7.86	11.03	3.90	6.44	3.27
4	6	4.97	8.24	11.17	4.69	8.37	2.81
5	0	5.72	8.29	11.06	4.49	6.39	2.64
5	1	5.31	6.93	11.14	2.71	3.97	4.69
5	2	5.74	7.31	11.00	4.70	5.76	2.53
5	3	5.13	8.22	10.91	4.61	6.26	2.95
5	4	5.92	8.25	11.05	5.01	8.65	2.25
5	5	5.81	7.88	11.12	3.45	7.76	3.80
5	6	5.52	8.50	11.03	4.11	7.30	3.36
6	0	7.09	7.99	11.04	5.47	7.13	2.19
6	1	6.56	6.57	10.95	4.27	6.16	3.69
6	2	6.25	7.39	11.05	3.33	7.25	3.51
6	3	6.29	8.26	11.01	4.43	6.65	2.27
6	4	8.32	8.29	11.08	4.69	9.05	3.11
6	5	7.04	10.94	11.13	4.40	6.98	3.35
6	6	6.60	6.61	11.04	4.18	6.84	3.26

Table 4.4: Jitter & Shimmer for artificial voiced speech with SNR = 10dB

believe that the best method is ‘waveform matching’ which is used by the Praat software suite and is essentially the basis of the ‘cross-correlation approach’. The cross-correlation technique is implemented in the software developed in this thesis and the values of HNR obtained have been compared, for artificial speech, with the values obtained from the Praat and MDVP software. The Table 4.5 summarises the comparison where thesis-HNR denote the cross-correlation method. There is no simulated shimmer or jitter, and the SNR is achieved by adding a zero-mean pseudo-random Gaussian sequence of appropriate variance to the purely periodic waveform.

$$s(n) = 8\sin(2\pi(200/F_s)n) + 6\cos(2\pi(400/F_s)n) \tag{4.18}$$

where the sampling rate F_s = 40000 Hz. This was an early experiment and all subsequent experiments use F_s= 44100 Hz. It should also be reported that the 32-bit Windows version 5.4.19 of the Praat software was used to produce Table 4.5, whereas all later experiments with ‘Praat’ use the 64-bit Windows version 6.0.04. Unfortunately these two versions of Praat give different results for exactly the same data files.

The ‘actual’ SNR in Table 4.5 is calculated from the actual noisy signal and differs slightly from the nominal SNR because of the limited number of samples. The Praat software is unable to calculate the HNR for SNR values much less that 0 dB, and though thesis-HNR continues to deliver values for this range, they are clearly not as accurate as those for higher values of SNR.

The results in Table 4.5 indicate that the thesis software produces HNR measurements close to the nominal values of SNR and those of Praat for a periodic test signal with varying degrees of added noise. The standard deviation of the difference between thesis-HNR and Praat-HNR over the SNR range -1 dB to 20 dB is 0.24 dB and the maximum difference is 0.53 dB in a measurement of 11 dB. The standard deviation of differences of thesis-HNR and Praat-HNR from the nominal values of SNR is 0.4086 and 0.3435 respectively.

Figure 4.10 represents Table 4.5 in graphical form. The occurrence of fundamental frequency halving (period doubling) is a strong possibility when estimating the fundamental frequency of noise affected signals.

This occurred in the Thesis software several times during the generation of Table 4.5. It could not be ascertained whether it also occurred with the Praat software. It may be inferred from the derivation of Equation (4.13) for HNR that for a strongly periodic signal, fundamental period doubling should have little effect on thesis-HNR

SNRdB	actual	thesis-HNR-dB	Praat-HNR-dB
20.00	20.01	20.25	20.38
19.00	19.01	19.07	18.95
18.00	17.95	18.36	18.07
17.00	17.04	17.75	17.59
16.00	16.21	16.12	15.97
15.00	14.91	14.72	14.77
14.00	14.15	14.35	14.41
13.00	13.18	13.75	13.24
12.00	11.97	11.77	11.95
11.00	11.34	12.09	11.56
10.00	10.42	10.13	10.27
9.00	9.26	9.51	9.35
8.00	8.07	7.75	8.02
7.00	7.31	7.85	7.87
6.00	5.79	6.96	6.49
5.00	4.83	5.46	5.18
4.00	4.42	4.72	4.51
3.00	2.90	3.22	3.43
2.00	1.74	2.35	2.54
1.00	1.24	1.80	1.89
0.00	-0.44	0.52	0.33
-1.00	-1.04	0.03	0.18
-2.00	-2.00	-1.45	undef
-3.00	-2.94	-1.75	unde

Table 4.5: Comparison of HNR measurements for periodic waveform (4.18) with added noise

since the signal will remain strongly periodic at twice its fundamental period. However, the noise averaging will now take place over twice as many samples, and thus be a little more accurate. Underestimating the period, for example by mistaking vocal tract resonance for the effect of vocal cords will affect the thesis-HNR though not catastrophically. A mistaken resonance must have a cross-correlation coefficient higher than that produced by the vocal cord periodicity and taking this as the fundamental periodicity will simply raise the estimated harmonic component slightly and produce a slightly less accurate noise estimate.

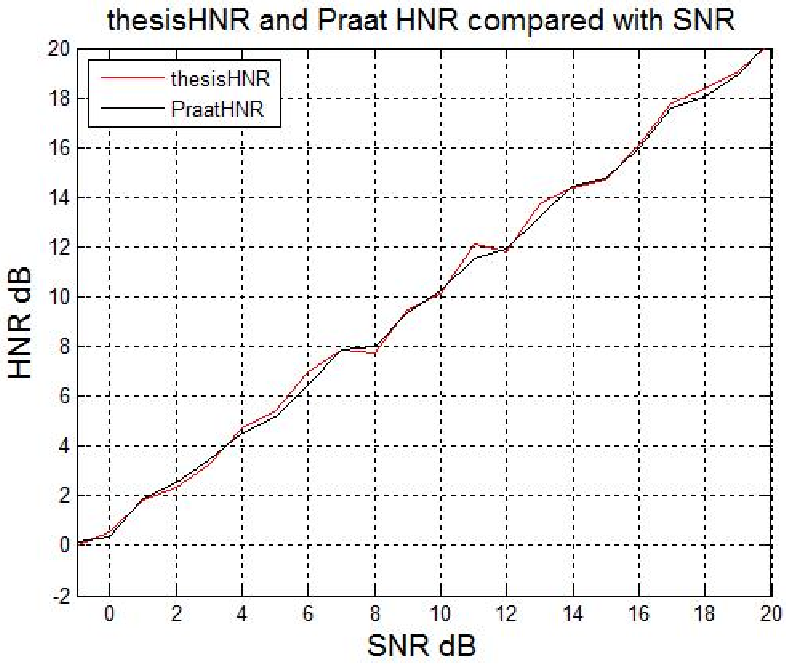


Figure 4.10: thesis-HNR & Praat-HNR compared with SNR for periodic waveform for thesis software and Praat software

4.9 Measuring HNR for artificial voiced speech

To further evaluate the 'thesis software' for HNR estimation and to compare it with the well known Praat and MDVP software suites, samples of artificial voiced speech were produced by exciting an all-pole vocal tract model, with glottal pulse shaping and lip-radiation filtering, by a periodic series of discrete time impulses. Again, the poles were given radii of 0.992, 0.99, 0.988, and 0.986 with frequencies ±610, ±1300, ±2450 and ±3600 Hz respectively to emulate the phoneme /a/. The sampling rate was 44.1 kHz. Pseudo-random Gaussian white noise of zero mean and appropriate variance was added to achieve signal-to- noise ratios ranging from -4 dB to 20 dB. In this experiment there was no simulated jitter or shimmer. Table 4.6, as presented graphically in Figure 4.11, indicates that there is a constant 1 dB discrepancy between the 'thesis' and Praat measurements of HNR for synthesised SNR values over the range -2 dB to 20 dB.

The Praat measurements are remarkably close to the nominal values of 'signal-to-noise' ratio (SNR) until it approaches 0 dB. If the constant 1 dB discrepancy is disregarded, the thesis software is also strongly indicative of the SNR value. The MDVP software gives 'noise to harmonic ratio' rather than HNR, but this was converted to

Synth-SNR	thesis-HNR	Praat	MDVP
20	21.74	20.15	9.20
19	20.35	19.03	9.20
18	19.38	18.09	9.20
17	18.32	17.20	9.20
16	17.38	16.19	8.86
15	16.29	15.15	8.86
14	15.11	14.04	8.53
13	14.20	13.17	8.23
12	13.22	12.17	8.23
11	12.21	11.19	7.95
10	11.32	10.23	7.95
9	10.15	9.09	7.21
8	9.18	8.19	7.21
7	8.31	7.19	6.77
6	7.22	6.20	6.38
5	6.33	5.24	6.19
4	5.46	4.40	5.68
3	4.42	3.36	5.37
2	3.39	2.29	4.94
1	2.58	1.47	4.55
0	1.71	0.50	4.08
-1	0.75	-0.43	3.90
-2	-0.27	undef	3.46
-3	-1.16	undef	3.01
-4	-1.63	undef	2.75

Table 4.6: Comparison of HNR measurements for artificial voiced speech with added noise

HNR by taking the reciprocal and then expressing the result in dB. The MDVP measurements are very different from the thesis-HNR and Praat-HNR measurements.

Although these results indicate that the Praat-HNR measurements should be preferred to the thesis-HNR measurements, we need to invoke the DSP software from the voice assessment software which is difficult with the scripting facilities provided by current versions (e,g. 6.0.04) of Praat. Clearly the constant 1 dB discrepancy needs investigation, but indications are that adopting the thesis software for HNR measurement appears reasonable.

4.10 Measuring other features

The software developed by this thesis derives the measurements listed in Table 4.7 from recordings of sustained vowels. The beginning and end of each sustained vowel

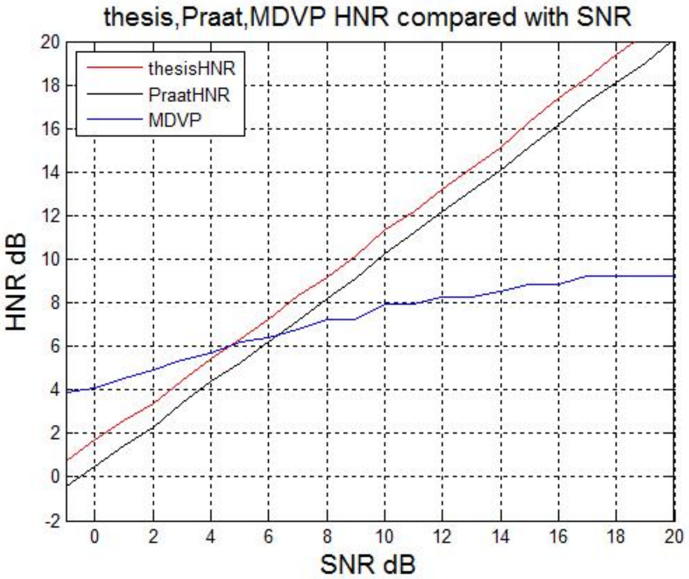


Figure 4.11: Comparison of HNR measurements for artificial voiced speech with added noise for thesis software, Praat and MDVP

are trimmed to remove silence and each sustained vowel is divided into a series of non-overlapping 22.676 ms (1000 sample) frames sampled at 44.1 kHz. The software also measures a fundamental frequency F_0 and a voiced/unvoiced decision for each frame. Where the thesis software is applied to connected speech, some features, such as HNR, jitter, shimmer, will be appropriate only for frames with pseudo-periodicity

The mean energy per frame (MEPF), the ratio of minimum to maximum energy per frame energy (RMMEPF) and the standard deviation of the frame-by-frame energy (STD-EPF) are easily computed for sustained vowels, and for vowels within connected speech. The MEPF of each vowel is normalized by dividing by the average of the MEPF values obtained for all ‘normal’ voices out of the 102 examples.

The mean ‘low-to-high spectral (L/H)’ ratio over the bandwidth 0 to 3 kHz with cut-off frequency 1.5 kHz is calculated for just voiced frames by two methods: digital filtering and frame-to-frame FFT spectral analysis with averaging. In principle both methods should give the same result. The standard deviation of the frame-to-frame measurements of this parameter are is also a measurement provided by the thesis software. The bandwidth (3 kHz) and the cut-off frequency (1.5 kHz) were chosen to highlight the damping of higher frequency energy in vowels that helps to characterize asthenia and other GRBAS components.

Feature Label	Feature	Definition
F1	API	Aperiodicity Index
F2	HNR	Harmonic to Noise Ratio
F3	Jitter	RAP jitter
F4	Shimmer	RAP shimmer
F5	MEPF	Mean Energy per frame
F6	RMMEPF	Ratio of minimum to maximum energy per frame energy
F7	STD EPF	Standard deviation of the frame-by-frame energy
F8	M-L/H	Mean ratio of low to high freq energy with c/o 1.5 kHz
F9	STD-L/H	Standard deviation of L/H spectral ratio with c/o 1.5 kHz
F10	Min /Max-L/H	Ratio of Max L/H-SR to min L/H SR (c/o 1.5 kHz)

Table 4.7: Features measured by thesis software

4.11 Feature measurement by ADSV

No single feature can completely characterise any of the GRBAS parameters for all voice disorders. Most published work uses a combination of different measures. In addition to the features listed in Table 4.8, there are others that the thesis software does not derive currently, but may also be useful. KayPentx [Kay96] have produced a multi-parameter measurement tool, the 'Analysis of Dysphonia in Speech and Voice' (ADSV) which was described in Chapter 2 Section 2.8. It is claimed that ADSV is not as dependent on pitch-cycle determination [AR09, ARJ$^+$10] as the Praat and MDVP software, though the exact details of the ADSV algorithms are not available. It produces measurements not provided by other packages that are applicable to both sustained vowels and connected speech. It was decided to augment the measurements obtained from the thesis software by some of the ADSV measurements in order to assess whether they are worth including in further developments of the thesis software. Table (4.8) lists some of the features that are measured by ADSV.

Feature measurement by ADSV proceeds as follows:

1. The beginnings and ends of sustained vowels and connected speech are trimmed to remove unwanted periods of silence.

Feature Label	Feature	Definition
F11	CPP	Cepstral Peak Prominence
F12	CPP STD	Std dev of CPP
F13	CPP Max	Max CPP for voiced frames
F14	CPP Min	Min CPP for voiced frames
F15	ML/H	Mean ratio of signal energy below 4 kHz to that above 4 kHz
F16	STD L/H	Std-dev of ML/H
F17	Max L/H	Max L/H spectral ratio (c/o 4 kHz) for voiced frames
F18	Min L/H	Min L/H spectral ratio (c/o 4 kHz) for voiced frames
F19	Mean CPP f_0 STD	Std-dev of the freqs of the cepstral peaks (60 Hz to 300 Hz) for voiced frames
F20	CSID	Cepstral/Spectral Index of Dysphonia

Table 4.8: Features measured by ADSV.

2. The speech, sampled at F_s = 44.1 kHz, is divided into a series of 1024 point overlapping frames with 75% overlap [AR09, ARJ$^+$10].

3. Voiced frames are detected by a voiced/unvoiced decision

4. For each analysis frame, the discrete Fourier transformation (DFT) is applied to a Hamming windowed version of the signal to obtain a power spectrum. The logarithm of this power spectrum, symmetrical about $F_s/2$, is then inverse-DFT transformed to obtain a real valued cepstrum as described by Baken [BO00]. From such a cepstrum, the cepstral peak prominence (CPP) may be derived as described earlier in this Chapter and recommended by [HCE94, HH96].

5. A combination of time and ‘cepstral time’ (quefrency) averaging is used to smooth the cepstrum prior to identification of the CPP [HH96]. A 7-frame cepstral averaging is carried out, with each smoothed cepstrum being calculated from the average of the current cepstrum, those from the three previous frames and those from the three subsequent frames. Cepstral averaging across time is followed by 11-bin quefrency averaging, in which each cepstral coefficient is replaced by the average of the current coefficient with the corresponding five previous and five subsequent cepstral coefficients.

6. ADSV measures several voice features from each frame. A ratio of low/high

frequency (L/H) spectral energy is calculated as a measure of 'spectral tilt' from the original unsmoothed window, (referred to as the DFT Ratio (DFTR) in [AR05, AR06]. This low/high spectral ratio has a cut-off frequency of 4 kHz rather than the 1.5 kHz used by the thesis software, and it takes the whole speech band-width into account. It is a different measurement.

7. The means and standard deviations of the L/H spectral ratio (c/o 4 kHz) and CPP are calculated for the entire signal. Standard deviations are collected because the various spectral/cepstral measures may be averaged across relatively long duration samples of non-stationary connected speech with vowel-consonant transitions and intonation. Previous studies have indicated that measures of variability may be effective in characterizing the severity of voice degradation [AR06, AR09, CKRT99, WS87].

8. The ADSV program defines a measure the Cepstral/Spectral Index of Dysphonia (CSID), which provides an assessment of dysphonic severity that can be used to measure a patients voice quality over time and before and after therapy or other intervention [ADS]. The CSID is calculated from a multiple regression formula derived from the correlation of results from the ADSV analysis with perceptual analysis of of trained scorers. The CSID provides an estimation of dysphonia severity which approximates a 100-pt. visual analog scale similar to the CAPE-V tool. Perceptual ratings of dysphonia severity were compared to acoustically-derived severity estimates using a multiple linear regression model. For estimating the severity in connected speech the model consisting of the cepstral peak prominence (CPP), CPP STD, the ratio of low-to-high spectral energy, and its standard deviation are the strongest contributors. A five factor CSID model incorporating all mentioned 4 acoustic features as well as gender was used to estimate severity in sustained vowel samples. Results showed strong relationships between perceptual and acoustic estimates in dysphonia severity in connected speech (r =0.81) and sustained vowels (r = 0.96). These correlation values were obtained for overall severity [ARJ$^+$10].

   The CSID formulas were calculated based on the default settings, and from data sampled at 25,000 Hz. The CSID has not been validated with alternative sampling rates; however, results from 22,050 Hz will likely be very similar to those obtained using the 25,000 Hz sampling rate. In addition, if any of the settings in the Advanced Options dialog are changed, the CSID is no longer applicable.

4.12 Identifying features likely to be indicative of GRBAS components.

4.12.1 Grade

'Grade' is the perceived degree of hoarseness or abnormality. All features mentioned above are capable of detect voice abnormality and therefore are likely to be relevant to Grade prediction.

4.12.2 Roughness

Roughness results from aperiodic vocal fold vibration which generates random noise-like energy in the voice and thus changes the perceived vocal quality [CL06]. Wolfe [WFC95] found fundamental frequency variation (jitter), peak amplitude variation (shimmer) and fundamental frequency tremor to be the best predictors of roughness. In research conducted by [MFW95] roughness was found to be best predicted by measurements of harmonic-to-noise ratio (HNR).

4.12.3 Breathiness

Breathiness is often due to incomplete glottal closure during the 'closed' phases of the phonatory cycles [HH96]. It may be associated with inflammation, vocal misuse [Aro90] or more serious and long-term conditions (see Chapter 2). It is normally detected by the harmonic-to-noise ratio (HNR) though there are other measurements such as 'glottal excitation to noise ratio' (GENR) which may give more indication of how the breathy sound is generated. GENR looks for correlation between the instantaneous energy of the breathiness and the different phases of vocal cord activity within each cycle.

There is evidence that the physiological effects of aging may include breathy voice [Hol87, RB74]. Klatt [KK90] have suggested that the presence of aspiration noise is the primary sign of breathiness. There are also conflicting findings on the relationship between spectral tilt and breathiness, with some researchers [Hil88, KK90] proposing that spectral tilt plays little or no role in the perception of breathy voice and other researchers [FJP76, Kli82] proposing that breathiness is associated with greater amounts of higher frequency energy. Other studies in this research area have aimed to measure the relationship between breathiness, GRBAS scoring and measurements [BPG04] of a relatively large set of acoustic features. The measurements fall into two categories:

1. Measures of signal periodicity such as HNR, CPP and API

2. Measures of spectral tilt such as ‘low to high spectral ratio’

4.12.4 Asthenia

With Asthenia, the overall speech energy, and, especially, the higher frequency harmonics of F_0 are attenuated causing a lack of volume, ‘richness’ and brightness in the perceived sound. The lack of volume and the spectral damping may be detected from the ‘energy’ and low to high spectral ratio measurements included in Tables (4.7) (thesis software) and (4.8) (ADSV). Other measurements such as CPP, API, and HNR may also detect the change in harmonic structure that occurs due to asthenia.

4.12.5 Strain

Strain due to speaking, or trying to speak with abnormality functioning vocal cords is perhaps most subjective GRBAS component and the most variable in its effect. Features we have associated with strain are:

1. An abnormality high fundamental frequency (F_0)

2. Unnatural and constantly changing periodicity

3. Roughness in the higher frequency range of the speech

These features are measured by F_0, by detecting changes in F_0 which are much slower than those detected by jitter and by HNR or CPP (or both).

4.13 Conclusions

Features that affect voice quality and may indicate the presence of voice disorder are described in this Chapter. ‘Thesis software’ written to measure and quantify some of these features is described. The use of commercial software (ADSV) for measuring some other features is also described. The ‘thesis software’ is evaluated against corresponding algorithms in the Praat software package. The Praat software package was chosen for this validation exercise since its techniques are reasonably well published and have been evaluated by many authors. Published comparisons with other packages such as MDVP have revealed significant differences which, to a degree, have been explained by the authors of Praat. The Praat software has scripting, but, with the current version (5.4.19), it could not be used without some user interaction. Hence the need for the 'thesis software'. This software uses what we believe to be the best available techniques, though its optimization to commercial levels of reliability, a great art in speech processing, is beyond the scope of this thesis. Voice has multidimensional properties and measurements of a single feature, we believe, can not quantity even one of the GRBAS components over a wide variety of voice conditions. Our aim in next chapter is to determine to what extent the features identified in this chapter as likely to be useful for predicting GRBAS components are actually useful. Some measurements, such as Jitter and Shimmer are made only for vowels but others will be made from voiced and unvoiced sections of connected speech.

There is clearly overlap in the information given by each measurement, and two or more measurements could be affected by the same voice feature. The measurements are not expected to be orthogonal, and, indeed, there is much to be said for having as many measurements as possible. Some measurements may concur for some voice conditions and diverge for others. Two measurements may diverge only for special conditions they are tuned to detect. These issues will also be considered in the next Chapter.

Chapter 5

Objective Prediction of GRBAS Scores

This chapter discusses the methodology and machine learning techniques that are used in this thesis for the objective prediction of GRBAS scores. The machine learning techniques will be trained using ‘training data’ consisting of examples of measured ‘voice features’, together with ‘reference’ GRBAS scores which can be considered ‘reliable’. We investigate the use of machine learning algorithms with the measurements of voice features that were discussed in Chapter 4.

The work in this chapter also considers feature dimensionality and its effect on the performance of the machine learning algorithms. A large set of voice measurements may be used to provide data for the machine learning and prediction algorithms. However, it may be feasible and preferred to take a smaller number of voice features for predicting the GRBAS scores. The literature for feature dimensionality reduction by introducing ‘feature extraction’ and ‘feature selection’ methods is reviewed and the two main methods for feature selection, ‘filters’ and ‘wrappers’ are explored. How the most relevant set of features for GRBAS prediction can be identified using feature selection methods is explained. With feature selection, the methods that appear likely to give the lowest prediction error are considered as appropriate methods. The results obtained from different models for the objective prediction of GRBAS scores are compared and analysed ‘with feature selection’ and ‘without feature selection’. The best feature subset for predicting each GRBAS component objectively is identified amongst different subsets. In this chapter L will denote the number of sample in the training set and is normally equall to 80.

5.1 Supervised learning prediction models

As mentioned in earlier chapters, regression models are preferred to classification models for GRBAS prediction. GRBAS scores can be considered as either 'quantitative' or 'categorical data'. Quantitative data have numerical values. Categorical data consist of elements which are non-numerical and members of different classes, or categories. When considered as categorical data, GRBAS scores are 'normal', 'mild impairment', 'moderate impairment' and 'severe impairment'. Dealing with categorical data is often referred to as a classification problem. Assigning the number '0' to 'normal', '1' to 'mild impairment', '2' to 'moderate impairment' and '3' to 'severe impairment' allows the GRBAS scores to be considered as quantitative data. In this thesis, 'GRBAS scores' are considered quantitative so regression models can be used. Regression problems take into account the numerical differences between the scores. Two supervised learning models are used for predicting the GRBAS scores. These models are multiple linear regression (MLR) and 'K-nearest-neighbour- regression' (KNNR) [BF85, Jia02]. A description of each prediction algorithm is given in the following sections. The term 'supervised' was explained in Section 2.13.2.

5.1.1 'Multiple Linear Regression' (MLR)

MLR is an approach for supervised learning problems. Equation (5.1) represents a linear regression model for predicting a variable Y with a dependency on the k predictor features X_1, X_2, ..., X_k. In our application, Y is the 'reference' or 'gold standard' obtained from the five scores and X_1, X_2, ..., X_k are the voice feature measurements discussed in Chapter 4. The constants $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients and ε is the prediction error which must be small in magnitude if the model is to be considered accurate. When $k = 1$ the model is simple linear regression and when $k>1$, it is multiple linear regression (MLR). The relationship between each individual predictor feature and the response Y may be estimated and modeled by a one-dimensional linear regression equation. This pre-supposes that the relationship is really linear or approximately so. However, instead of producing a separate simple linear regression model for each voice feature, a better approach is to extend the simple linear regression model to Equation (5.1) so that it can directly accommodate multiple predictors.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots.. + \beta_k X_k + \varepsilon \tag{5.1}$$

Assume that Y represents the 'grade' dimension of GRBAS and that we are given values of 'grade' with corresponding measurements of voice features such as HNR, API, Jitter, shimmer, etc for a large number (N) of subjects. Multiple linear regression can be applied to model the relationship between the 'grade' scores and these voice features. Having calculated the beta parameters for this relationship, the grade score for a new subject can then be predicted from measurements of the specified voice features in the new subject.

When there are N subjects, we have N values of Y, called $Y_1, Y_2, ...Y_\mathrm{N}$. For each Y_i, we have corresponding values of the k voice features which we call $X_{\mathrm{i}1}$, $X_{\mathrm{i}2}$,.....X_{ik}. If the MLR model represented by equation (5.1) is likely to be useful, the following set of N equations should apply with suitably small values of $\varepsilon_{1}, \varepsilon_{2}, \varepsilon_{3}, ...\varepsilon_{N}$

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + ... + \beta_k X_{1k} + \varepsilon_1 \tag{5.2}$$

$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + ... + \beta_k X_{2k} + \varepsilon_2 \tag{5.3}$$

$$Y_\mathrm{n} = \beta_0 + \beta_1 X_{\mathrm{n}1} + \beta_2 X_{\mathrm{n}2} + ... + \beta_k X_{nk} + \varepsilon_3 \tag{5.4}$$

These N equations can be represented in matrix form as

$$\underline{Y} = X.\underline{\beta} + \underline{\varepsilon} \tag{5.5}$$

The accuracy of the model is determined by the magnitudes of $\varepsilon_1, \varepsilon_2, ...\varepsilon_N$. A measure of this accuracy is the sum of squares:

$$E = {\varepsilon_1}^2 + {\varepsilon_2}^2 + ... + {\varepsilon_N}^2 = \underline{\varepsilon}^T \underline{\varepsilon} \tag{5.6}$$

The smaller the value of E, the more accurate is the MLR model in representing the known data. It is pre-supposed that the more accurate the model is for known data, the more accurate it is likely to be for unknown data (new subjects). The next section considers how the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, i.e. the elements of vector $\underline{\beta}$, may be calculated such that E is minimised for the known data.

5.1.1.1 Estimating the regression coefficients

The N by (k+1) matrix X in equation (5.5), whose first column contains just ones is referred to as the design matrix. Vector $\underline{\beta}$ contains all the regression coefficients

$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. It may be shown [Re92] that the vector of regression coefficients, $\underline{\beta}$, that minimises $E = \underline{\varepsilon}^T \underline{\varepsilon}$ is:

$$\underline{\hat{\beta}} = X^{\#}.\underline{Y} \tag{5.7}$$

where $X^{\#}$ is the 'pseudo-inverse' of the non-square matrix X and is equal to:

$$X^{\#} = (X^T X)^{-1} X^T \tag{5.8}$$

and where 'T' denote 'matrix- transpose'. This assumes that, there are enough values of Y to make $X^T X$ non-singular. In fact, the MATLAB function 'pinv' was used to produce pseudo-inverse' $X^{\#}$. Having obtained values of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ in vector $\hat{\beta}$, the multiple linear regression model:

$$\hat{Y} = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots.. + \hat{\beta}_k X_k \tag{5.9}$$

produces, for a set of feature measurements $X_1, X_2, \ldots. X_\mathrm{k}$ from a given subject, an estimate $\hat{Y}$. If the MLR model is successful, $\hat{Y}$ will be close to Y for all the subjects in our data-base, and may then be expected to be close to a GRBAS score for an unknown subject for which feature measurements $X_1, X_2, \ldots. X_\mathrm{k}$ have been made. The calculation of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ as outlined above is one form of 'machine learning'. The information that has been learned is held in these k (small k) coefficients.

5.1.2 'K Nearest Neighbour' (KNN) Regression

KNN is a non-parametric method for classification and regression [Jia02]. It does not make any assumption about the distribution of the data or whether the relationship between the feature measurements and the target scores is likely to be linear or non-linear. With KNN, the 'machine learning' information is held in the data-base itself, so just giving the machine a data-base is all that is required of the learning process.

The data-base should consist of the k (small k) feature measurements and the 'reliable' reference scores for each of the N subjects. K (large K) is an integer parameter that defines the way the KNN approach predicts a score for a new subject from measurements of its k parameters. The prediction is based on the known scores for K members of the database which are chosen according to the distance of their measured features from the measured features of the new subject. The concept of 'distance' can be defined in various ways: for example, Manhatten distance (sum of moduli [RMR07]) or Euclidian distance. We have taken the Euclidean distance which means the square-root of the sum of the squared differences between the k feature measurements of the new subject and the corresponding feature measurements of the data-base subject.

When KNN is used for classification or regression, the 'distance' between the new subject and each of the N data-base subjects is calculated. Then K data-base subjects are selected as being those that are nearest to the 'new subject' according to their feature measurements and the Euclidean distance measure. This is out of all possible choices of K data-base subjects. To take an example, assume that K= 5, and that the 5 nearest neighbours to a new subject are as shown in Table 5.1.

subject	distance	classification	label
7	0.1	'moderate'	2
15	0.4	'normal'	0
22	0.2	'severe'	3
25	0.8	'normal'	0
27	0.1	'mild'	1

Table 5.1: Illustration of five nearest neighbours

If KNN is used for 'classification', majority voting could be used to select 'normal' as the classification result in this illustration, since 'normal' occurs more often than the other classifications of 'grade'. Using '0' as a label for 'normal', '1' for 'mild impairment', and so on does not change this classification process if we continue to regard 0, 1, 2 and 3 as classification labels and not numbers. The classification result would still be '0. However, KNN regression (KNNR) would give a different result by considering the magnitudes of the numbers to be measures of the degree of impairment.

A simple form of KNN regression takes the arithmetic mean (average) of the scores of the K nearest neighbours as the result. In the example, this would be the average of 0, 0, 1, 2, 3 which is 1.25. A preferred form of KNN regression takes a weighted average of the scores of the K nearest neighbours where each score is weighted according to the 'near-ness' of the data-base subject to the new subject. Defining nearness as the reciprocal of the Euclidean distance with provision for accommodating a distance of zero, allows a weighted arithmetic mean, defined as follows, to be used to predict the required new score.

$$prediction = \frac{n_1 \times s_1 + n_2 \times s_2 + n_3 \times s_3 + n_4 \times s_4 + n_5 \times s_5}{n_1 + n_2 + n_3 + n_4 + n_5} \tag{5.10}$$

where n_1, n_2, n_3, n_4, and n_5 are the 'nearness' factors and s_1, s_2, s_3, s_4, s_5 are the reference scores. For the illustration in Table (5.1), the weighted prediction would be: $(2\times 10+0+5\times 3+1.25\times 0+10\times 1)/(10+2.5+5+1.25+10) = 45/28.25 = 1.59$ This may subsequently be rounded to 2.

If all neighbours are equidistant from the new subject, the weighted average becomes the normal average. If the new subject happens to coincide with a data-base subject, having identical feature measurements and therefore a distance of zero, a large nearness factor would be generated probably ensuring that the score of the new subject would be identical to the coinciding data-base subject. Taking a non-weighted average would not necessarily ensure this. However, it is not always desirable especially if there is some unreliability in the feature measurements or the reference scores. Weighted KNN is regression, well researched [LC11, PIGP+93] though it has issues that would be suitable topics for further research. A clear advantage over the simple approach is that the value of K becomes less critical.

Regression takes into account that there is a stronger difference in severity between scores 1 and 3 than between 1 and 2. Majority voting classification would not take this difference into account.

5.1.2.1 Feature scaling with KNNR

The numerical values of the feature measurements for each subject in the data-base can have widely different dynamic ranges. The ranges can differ by many orders of magnitude such as 0 to about 5% for jitter and 0-100 for CSID. This dynamic range feature variability can adversely affect the result of KNN classification or regression since large numerical differences in feature measurements with large dynamic ranges may dominate into insignificance small numerical differences in feature measurements with small dynamic ranges. To improve the performance of KNN regression, all feature measurements are scaled as in Equation (5.11) to make the mean of each feature equal to zero and the standard deviation equal to 1.

$$f_{\mathrm{m\_scaled(n,i)}} = (f_{\mathrm{m\ (n,i)}} - M(i))/\sigma(i) \quad for \quad i = 1,2,....,k \tag{5.11}$$

where $f_{\mathrm{m(n,i)}}$ is the ith feature-measurements and $f_{\mathrm{m\_scaled\ (n,i)}}$ is the scaled version of $fm(n,i)$ for each subject $n = 1,2,\ldots,N$. For each feature measurement i, M(i) and $\sigma(i)$ are the mean and standard deviation of the unscaled measurements $f_{\mathrm{m(n,i)}}$ over all subjects $n = 1,2,\ldots,N$.

Feature scaling is not necessary for MLR because the beta coefficients in Equation

(5.1) effectively weight the contribution of each measured feature. Small values of beta are normally obtained for feature measurements with large dynamic range, and vice-versa.

5.1.3 Performance of prediction models

An important task is to decide for the collected data-base which prediction model (KNNR or MLR) can produce the best results when the parameters, such as K, have been chosen optimally. Ultimately, it must be decided whether either of these models can be the basis of a useful voice quality assessment system. To compare the models, we need have some way to measure the extent to which predictions to GRBAS scores obtained from the models actually match the scores obtained from our five SLT scorers. In applications of regression, a commonly used measure of prediction accuracy is the root-mean-squared-error (RMSE). The RMSE between 'predicted' and 'reference' values of GRBAS parameters computed over all subjects in the data-base as some indication of whether the objective scoring technique is likely to work with unseen data that was not used to fit the model. This value of RMSE is referred to as the 'training-RMSE' Although the 'training-RMSE' can give us some confidence in our training methodology, this may be false confidence as we cannot be certain that it will accurately predict the performance of the system for future subjects not in the data-base. If the training-RMSE is poor meaning that the system is not working well for the data it was trained on, this strongly suggests that something is wrong with the system. We are not primarily interested in the performance of the system for subjects used to train the model, since we already know the GRBAS scores for these subjects.

We may set out to produce a model that gives the lowest possible 'training-RMSE' over the whole database of N subjects. However, we would like to be able to test the performance of the model afterwards using unseen data not in the data-base. But we may not have any more data. Therefore, it is necessary to set aside some of the subjects in our data-base for testing. Assume we use a sub-set of L subjects for training. The 'training-RMSE' between the predicted $(\hat{Y}_{\mathrm{i}})$ and the observed (or reference) values (Y_{i}) of some GRBAS parameter, for L subjects, is:

$$TrainingRMSE = \sqrt{(1/L)\sum_{i=1}^{L}(\hat{Y}_{\mathrm{i}} - Y_{\mathrm{i}})^2} \tag{5.12}$$

The training-RMSE will be small if the predicted GRBAS scores are very close to the reference GRBAS scores for these L training subjects and will be larger if some of them differ substantially. The 'training Normalised Root-Mean-Square-Error' (training-NRMSE) is defined as the following percentage:

$$\begin{aligned} TrainingRMSE(\%) &= (TrainingRMSE/(GRBAS_{\max} - GRBAS_{\min})) \times 100 \\ &= (TrainingRMSE/3) \times 100 \end{aligned} \tag{5.13}$$

This is a percentage of the maximum GRBAS score of '3'.

5.2 Training and testing

In our experiments, there were k =20 voice feature measurements per subject. Ten of these were produced by the 'thesis software' described in Chapter 4, and the other ten were obtained using the Kay-Pentax ADSV software. The features and their acronyms are presented in Tables (4.7) and (4.8). To test the capabilities of the MLR and KNNR methods for predicting each GRBAS component [WF05], L= 80 subjects were chosen for training purposes from the data-base of N = 102 subject. The remaining (N-L) = 22 subjects were set aside for testing purposes.

The 80 subjects selected for the training process were divided into a number of 'folds' to allow 'cross-validation' [K$^+$95]. With ten folds, each containing 8 randomly chosen subjects, this meant that the model was trained using the 72 subjects contained in nine folds chosen at random out of the ten folds. The remaining one fold, containing 8 subjects, was then available to be used for a validation test (see later). This process, referred to as a 'trial', was repeated 19 times; each time with a different random choice of L subjects for the training subset and therefore a different testing sub-set. We therefore performed 20 trials, for each trial, there was a training phase and a testing phase.

The training phase for MLR calculates a set of beta coefficients from the L selected training subjects as outlined in Section 5.1.1. The training phase for KNNR just requires a database to be populated with the L selected training subjects, and an appropriate value of K to be selected as indicated in Section 5.2.1.

The testing phase was the same for both MLR and KNNR. The trained models were applied to the 22 testing subjects and the results obtained for each GRBAS component were compared with the known reference value for that component. A value of the 'testing-RMSE' and the 'testing Pearson's correlation coefficient' testing-corr was obtained over the 22 testing subjects where:

$$TestingRMSE = \sqrt{(1/(N-L))\sum_{i=1}^{N-L}(\hat{Y}'_{\mathrm{i}} - Y'_{\mathrm{i}})^2} \qquad (5.14)$$

$\hat{Y}'_{\mathrm{i}}$ are the predicted values of a GRBAS component and Y'_{i} are the reference values of that component for the N-L members of the testing subset. The testing RMSE is normalised as for the training RMSE in Equation (5.13) to produce a testing NRMSE expressed as a percentage. The coefficient 'testing-corr' for each trial is simply the Pearson correlation between the 22 values of $\hat{Y}'_{\mathrm{i}}$ and the corresponding reference values Y'_{i}. After conducting 20 trials, we have 20 values of testing-RMSE and 20 values of testing-corr for each GRBAS component. These are averaged over the 20 trials to obtain overall values of these parameters referred to as 'NRMSE' and 'Corr'. Figure 5.1 shows a full cross-validation procedure.

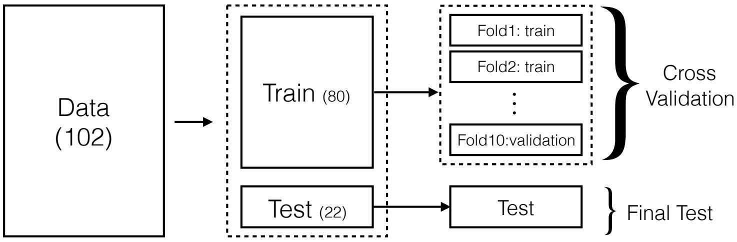


Figure 5.1: Full cross-validation procedure, with train, validation and test sets.

5.2.1 Optimum K for the GRBAS prediction by KNNR

With KNNR, the accuracy of the prediction will be affected by the value of K, which is the number of nearest neighbours chosen from the data-base. To determine an appropriate value of K for each 'trial', the 10 fold cross-validation approach mentioned above was used as now described.

Setting aside 'fold 1' for 'validation', the remaining 72 subjects, from the other nine folds, were used to populate the KNNR data-base. The KNNR procedure was then applied repeatedly to each of the 8 subjects in 'fold 1', firstly with K=1, then with K=2, K=3, and so on up to K=10. The normalised RMSE between predicted and reference GRBAS scores was computed over the 8 subjects for each value of K.

This procedure was repeated with 'fold 2' set aside and the remaining 72 subjects used to populate the KNNR database. The KNNR procedure was applied repeatedly to each subject in 'fold 2' to obtain a second normalised RMSE for each value of K. After repeating the procedure for ‘fold3’, ‘fold4 and so on up to ‘fold 8’ the eight values of normalised RMSE obtained were squared, averaged and square-rooted to obtain an overall ‘validation-NRMSE’ for each value of K for one trial.

This whole sequence of calculations was carried out 20 times, each time with a different random choice of L = 80 training subjects, to complete a series of 20 trials. This resulted in 20 sets of values of overall validation-NRMSE, one set for each trial. Each set contained a prediction error for each GRBAS component for each value of K in the range 1 to 10, i.e. 50 values. For each GRBAS component, for each value of K, the 20 values of validation-NRMSE, one for each trial, were squared, averaged and square-rooted, to obtain an trial-averaged validation-NRMSE, referred to as the ‘average NRMSE over 20 trials’. The term ‘validation’ is perhaps a misnomer since the aim of the sequence of calculations just described is to find a suitable value of K rather than to validate anything. The sequence of calculations produces the five graphs shown in Figure (5.2) for ‘grade’, Figure (5.3a) for ‘roughness’, Figure (5.3b) for ‘breathiness’, Figure (5.3c) for ‘asthenia’ and Figure(5.3d) for ‘Strain’. Each of these graphs shows the value of trial-averaged NRMSE for each value of K in the range 1 to 10.

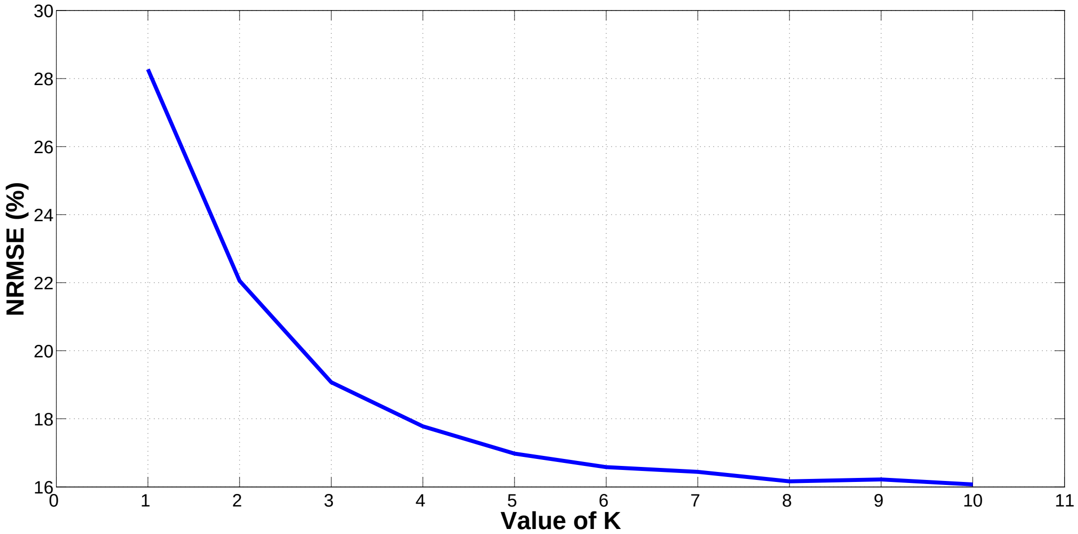
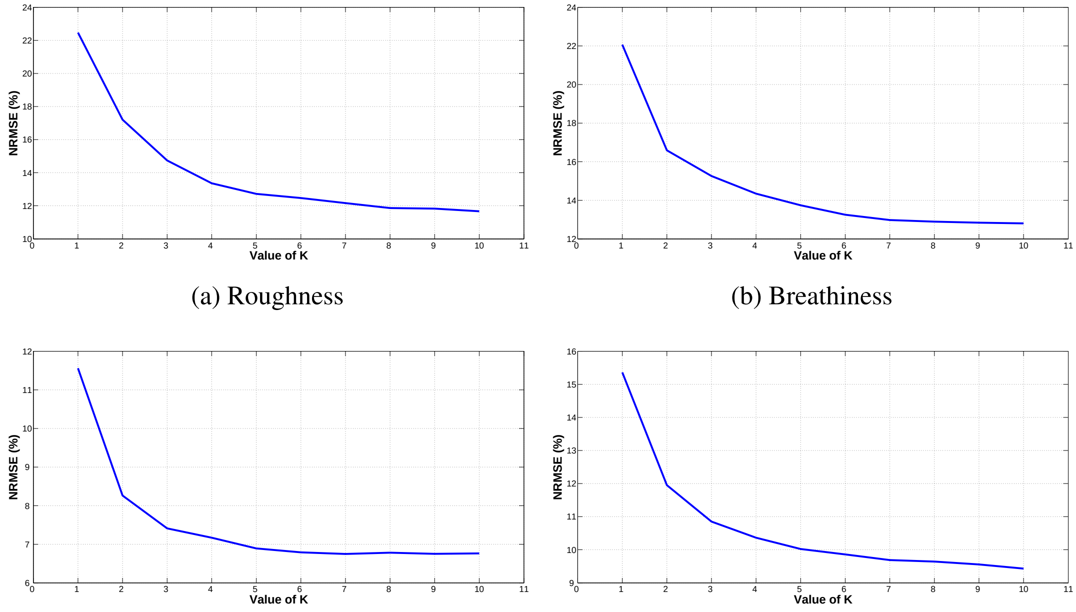


Figure 5.2: Average of Trial-Averaged-Validation-NRMSE for K=1 to K=10 (Grade)

As well as producing the averaged results in Figures (5.2) to (5.3d), the optimum value of K for each trial was obtained by observing the value of K that gives the lowest value of validation-NRMSE for each trial. The optimum value of K for ‘grade’ in each trial is shown in the histogram in (5.4). Similar histograms are produced for the other



(a) Roughness

(b) Breathiness

(c) Asthenia

(d) Strain

Figure 5.3: Average of Trial-Averaged-Validation-NRMSE for K=1 to K=10 (RBAS)

GRBAS components in Figures (5.5a) to (5.5d). The histograms show that the best value of K is not necessarily the largest. Figures (5.2) to (5.5d) were all produced for the case where all available 20 features are used for the predicting the GRBAS scores.

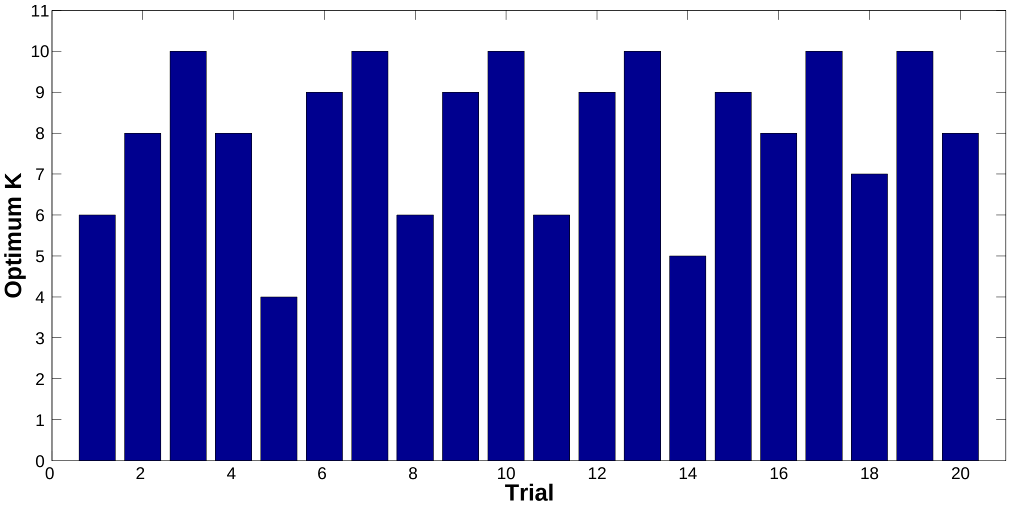
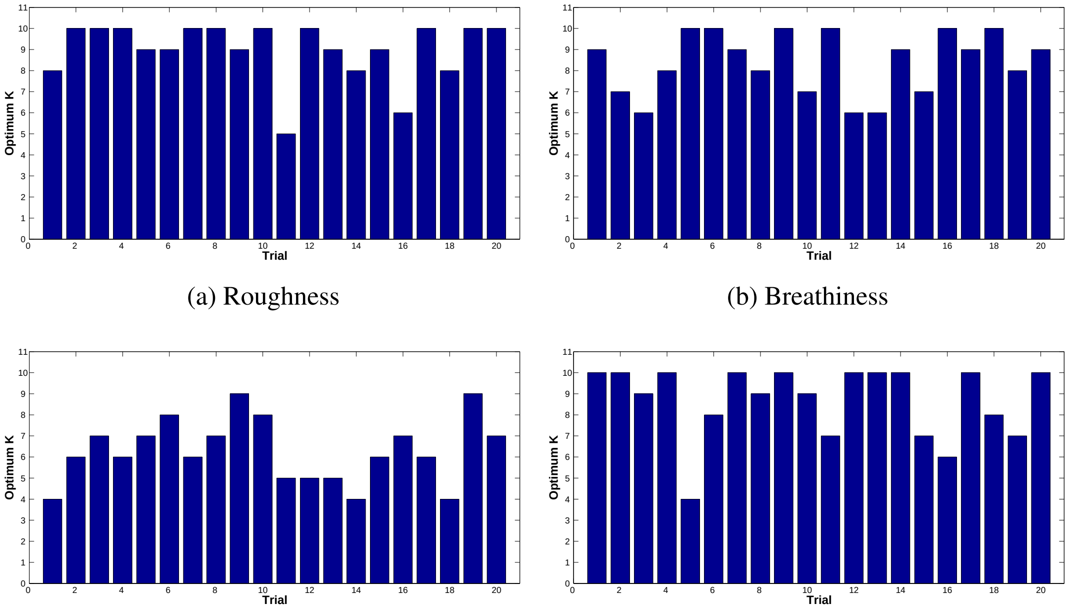


Figure 5.4: Optimum K for 20 features in each Trial (Grade)

Although there is clearly some variability in the optimal value of K for different training and validation data-sets, and for different GRBAS components, in practice, a fixed value of K must be adopted for a practical system. Figures (5.2) to (5.3d) show that the trial-averaged value of NRMSE does not change significantly for K greater than about 5 for any of the GRBAS components. Therefore, in subsequent experiments with KNNR, K=5 will be adopted. Finding an appropriate value of K completes the training phase for KNNR.



(a) Roughness

(b) Breathiness

(c) Asthenia

(d) Strain

Figure 5.5: Optimum K for 20 features in each Trial (RBAS).

5.3 Prediction by all features

The procedure for training and testing outlined in Section 5.2 was carried out for both the MLR and KNNR models applied to each of the five GRBAS components. For the KNNR testing, an appropriate value of K had to be chosen as addressed in the previous section. In each case, the training was repeated for 20 ‘trials’ where, for each trial, a new randomly selected subset of 80 subjects was used for training the model with the remaining 22 subjects used for testing it. Initially all 20 feature measurements were used. Table 5.2 compares the averaged testing-NRMSE and testing-corr values obtained for each GRBAS component using MLR and KNNR with all 20 features taken into account. The averaging is over all 20 trials.

It may be seen in Table 5.2 that the performances of these two techniques appears

	MLR		KNNR	
GRBAS	NRMSE%	Corr	NRMSE%	Corr
G	21.44	0.71	23.93	0.65
R	19.76	0.60	19.92	0.59
B	21.27	0.52	21.72	0.48
A	15.93	0.63	15.83	0.66
S	16.97	0.65	17.52	0.63

Table 5.2: Prediction results using all 20 features.

quite similar according to the 'NRMSE' and 'Corr' measurements and the testing procedure outlined earlier. The NRMSE measure for MLR is slightly lower (better) than that for KNNR for 'G', 'R', 'B' and 'S' and only very slightly higher for 'A'. This is consistent with the 'Corr' measure for MLR which is higher for 'G, 'R', 'B' and 'S' and lower for A. The significance of the values obtained as an indication of the practicality of an objective voice assessment system based on MLR or KNNR will be more fully addressed in the next Chapter. NRMSE percentages around 17% represent about half an increment in GRBAS scoring and, according to Table 3.1 in Chapter 3, Pearson correlation between 0.5 and 0.8 represents 'strong linear correlation'. Therefore, although our hope for NMRSE = 0 and corr = 1 have not been realised, the values in Table 5.2 do not appear totally discouraging. However it would be good to find ways of improving these values and the following sections address this issue.

5.3.1 Discussion

The 20 features cannot be expected to produce measurements which are statistically independent of each other. Some will be strongly correlated. Also the usefulness of each of these features as a predictor to a GRBAS score will be far from uniform. Some features may be strongly indicative and others may be less so. An indication of the usefulness of each measured feature for predicting a given GRBAS component can be obtained by computing the Pearson correlation between the reference scores and the corresponding measurements of that single feature. There is an assumption here that a linear approximation to the relationship between the feature and the GRBAS score gives a reasonable indication of the strength of the relationship, if there is one.

Where there are feature measurements which are strongly correlated with each other, the use of 'Principle Component Analysis' (PCA) or a related technique can clearly reduce the dimensionality of the prediction procedure and thus save much computation. However, such a reduction in dimensionality would not, in itself, improve the prediction results except as a means of identifying and eliminating linear combinations of features that are considered adequately represented by other linear combinations of features. The correlation of features or linear combinations of features with the prediction error is not taken into account with this process, though it can be in enhanced forms of PCA such as Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). We found that including the same strongly indicative feature twice does not affect the accuracy of the prediction, only the computation required. Therefore, we have not employed PCA in this thesis since we prefer to eliminate features individually on the basis of their direct effect on the output. In future work, it may be useful to investigate PCR, rather than PCA, for the feature selection procedures.

There is a reason to believe that including features that are not indicative of the GRBAS component being predicted may degrade the prediction, essentially by introducing noise or confusion. Therefore, we performed some experiments with feature reduction to try to improve our results by eliminating some features. Applying the feature selection procedures that will be introduced later in this section becomes computationally intensive when there are 20 features. With 20 features there are 2^{20} possible combinations of features. Therefore we started with a simple feature selection procedure for eliminating up to ten of the twenty features. We investigated the extent to which the prediction could be improved or degraded by discarding up to ten features from the original set of 20.

5.4 Progressively discarding features

The number of features was reduced to 19 by discarding the feature whose measurements had the lowest value of Pearson Correlation with the GRBAS component over 102 examples. The Pearson correlation values for ‘G’,’R’,’B’,‘A’ and ‘S’ are given in Tables (5.8) to (5.12) respectively. This procedure was repeated by discarding the feature with the next lowest correlation, and so on until only ten features remained. Tables (5.3), (5.4), (5.5), (5.6) and (5.7) show the values of NRMSE and Corr for the different numbers of features. As for table 5.2, these tables were computed for 22 randomly selected testing subjects and averaged over 20 trials.

These tables show that reducing the number of features from 20 to 10 does not

	MLR		KNNR	
Number of Features	NRMSE	Corr	NRMSE	Corr
20	21.44	0.71	23.93	0.65
19	21.12	0.72	23.46	0.67
18	21.65	0.71	23.50	0.66
17	21.72	0.70	23.69	0.66
16	22.04	0.70	23.73	0.66
15	21.63	0.70	23.88	0.66
14	21.83	0.70	23.93	0.65
13	21.65	0.70	24.17	0.65
12	22.06	0.69	24.06	0.65
11	22.15	0.69	24.01	0.65
10	21.82	0.70	24.14	0.65

Table 5.3: Prediction Results for Grade by MLR and KNNR

	MLR		KNNR	
Number of Features	NRMSE	Corr	NRMSE	Corr
20	19.76	0.60	19.92	0.59
19	19.58	0.60	19.77	0.60
18	19.34	0.62	19.61	0.61
17	19.60	0.61	19.77	0.60
16	19.55	0.61	19.35	0.61
15	19.04	0.62	19.46	0.61
14	18.78	0.63	19.41	0.61
13	18.57	0.63	19.40	0.61
12	18.23	0.65	19.70	0.60
11	18.11	0.65	19.66	0.59
10	17.97	0.66	20.12	0.58

Table 5.4: Prediction Results for Roughness by MLR and KNNR

significantly affect the accuracy of the KNNR prediction for any of the GRBAS components, though there is an improvement for MLR predicting ‘Roughness’, ‘Breathiness’ (small) and ‘Asthenia’. Both Pearson Correlation and NRMSE are consistent in their indications of the changing accuracy for each GRBAS component. For predicting ‘Grade’ by MLR, the NRMSE and correlation remain about 21.44% and 0.71. When applying KNNR for Grade prediction, the NRMSE remained about 2% higher and the correlation about 6% lower than MLR as the number of features were reduced. For ‘R’ the MLR and KNNR assessments of NRMSE and Corr started off similar, but MLR improved significantly (according to both assessments) as the number of features was reduced; KNNR did not change much. For ‘B’ and ‘S’, there was little change, though

	MLR		KNNR	
Number of Features	NRMSE	Corr	NRMSE	Corr
20	21.27	0.52	21.72	0.48
19	21.01	0.53	21.37	0.50
18	20.88	0.54	21.67	0.48
17	21.52	0.51	21.85	0.47
16	21.75	0.51	21.77	0.49
15	21.78	0.51	21.87	0.50
14	21.60	0.52	22.22	0.48
13	21.10	0.53	21.85	0.50
12	21.02	0.52	21.70	0.50
11	20.72	0.53	21.90	0.50
10	20.60	0.54	21.71	0.51

Table 5.5: Prediction Results for Breathiness by MLR and KNNR

	MLR		KNNR	
Number of Features	NRMSE	Corr	NRMSE	Corr
20	15.93	0.63	15.83	0.66
19	15.75	0.64	15.75	0.66
18	15.67	0.65	15.66	0.67
17	15.64	0.65	15.85	0.67
16	15.26	0.67	16.20	0.66
15	15.07	0.68	16.07	0.66
14	14.87	0.69	15.74	0.67
13	14.80	0.69	15.89	0.67
12	14.63	0.70	15.87	0.67
11	14.48	0.70	15.55	0.67
10	14.26	0.71	15.79	0.67

Table 5.6: Prediction Results for Asthenia by MLR and KNNR

for ‘A’ there was significant improvement in MLR (not KNNR) as the number of features were reduced. Overall, the highest correlation (0.72) was obtained for Grade with 19 features and Asthenia (0.71) with 10 features. The lowest NRMSE values (around 14%) were obtained for Asthenia prediction.

These tables appear to indicate that there may be some advantage, with respect to MLR only, in reducing the number of features by discarding those with lowest Pearson correlation to some of the GRBAS components. There is no evidence that any accuracy is lost by this action applied to either MLR or KNNR.

Tables 5.3 to 5.7 showed the ‘Pearson correlation coefficient’ and ‘testing-NRMSE’ between predicted and reference GRBAS scores, for testing examples of 22 subjects,

	MLR		KNNR	
Number of Features	NRMSE	Corr	NRMSE	Corr
20	16.97	0.65	17.52	0.63
19	16.56	0.66	17.40	0.64
18	16.83	0.65	17.48	0.63
17	16.80	0.65	17.51	0.62
16	16.89	0.64	17.37	0.63
15	17.28	0.63	17.46	0.63
14	17.45	0.62	17.56	0.63
13	17.12	0.64	17.66	0.62
12	16.92	0.65	17.38	0.64
11	16.78	0.65	17.70	0.63
10	16.68	0.66	17.76	0.62

Table 5.7: Prediction Results for Strain by MLR and KNNR

averaged over 20 trials. This was done for between 20 and 10 features selected according to the correlation values in Tables 5.8 to 5.12.

	Grade	
Feature Label	Corr	P-Value
F4	0.70	4.44E-16
F14	-0.64	6.79E-13
F20	0.62	4.50E-12
F11	-0.61	1.40E-11
F13	-0.48	5.20E-07
F1	0.47	9.48E-07
F5	-0.46	1.02E-06
F2	0.46	1.08E-06
F19	0.43	6.19E-06
F6	-0.43	7.78E-06
F12	0.42	1.33E-05
F18	-0.41	2.43E-05
F16	0.40	2.98E-05
F10	-0.39	5.63E-05
F7	-0.37	1.42E-04
F15	-0.31	1.38E-03
F3	0.23	2.10E-02
F9	0.21	3.09E-02
F17	-0.16	1.04E-01
F8	0.12	2.45E-01

Table 5.8: Correlation of individual feature measurements with Grade score.

Feature Label	Roughness	
	Corr	P-Value
F4	0.68	4.48E-15
F14	-0.59	8.69E-11
F20	0.57	4.82E-10
F12	0.52	2.89E-08
F11	-0.51	5.92E-08
F19	0.44	4.14E-06
F5	-0.43	7.28E-06
F1	0.38	1.11E-04
F7	-0.37	1.38E-04
F2	0.36	1.78E-04
F18	-0.36	2.33E-04
F16	0.34	5.74E-04
F6	-0.32	9.62E-04
F13	-0.30	2.27E-03
F10	-0.26	8.61E-03
F15	-0.25	1.06E-02
F3	0.23	2.07E-02
F9	0.17	9.91E-02
F17	-0.14	1.51E-01
F8	0.10	3.35E-01

Table 5.9: Correlation of individual feature measurements with Roughness score.

Feature Label	Breathiness	
	Corr	P-Value
F13	-0.54	3.93E-09
F11	-0.54	6.00E-09
F4	0.50	8.23E-08
F20	0.48	3.58E-07
F14	-0.47	5.32E-07
F5	-0.46	1.50E-06
F2	0.43	8.95E-06
F1	0.42	1.23E-05
F7	-0.37	1.21E-04
F10	-0.36	1.77E-04
F18	-0.36	1.89E-04
F19	0.32	1.00E-03
F15	-0.32	1.19E-03
F16	0.31	1.40E-03
F6	-0.28	4.50E-03
F17	-0.20	5.04E-02
F3	0.16	9.98E-02
F12	0.14	1.75E-01
F9	0.06	5.58E-01
F8	-0.04	7.06E-01

Table 5.10: Correlation of individual feature measurements with Breathiness score.

Feature Label	Asthenia	
	Corr	P-Value
F4	0.66	3.67E-14
F11	-0.64	6.30E-13
F20	0.63	1.09E-12
F14	-0.63	1.14E-12
F13	-0.56	8.08E-10
F19	0.48	4.40E-07
F5	-0.47	5.31E-07
F16	0.46	1.59E-06
F18	-0.45	1.85E-06
F1	0.45	1.89E-06
F10	-0.45	2.20E-06
F2	0.45	2.50E-06
F6	-0.45	3.08E-06
F7	-0.35	2.96E-04
F15	-0.35	3.25E-04
F12	0.33	6.49E-04
F9	0.23	2.12E-02
F17	-0.18	7.89E-02
F3	0.15	1.23E-01
F8	0.12	2.15E-01

Table 5.11: Correlation of individual feature measurements with Asthenia score.

Feature Label	Strain	
	Corr	P-Value
F4	0.69	1.82E-15
F14	-0.61	9.30E-12
F20	0.59	6.02E-11
F11	-0.58	3.02E-10
F6	-0.47	5.82E-07
F13	-0.45	2.47E-06
F18	-0.42	1.04E-05
F5	-0.42	1.46E-05
F19	0.41	1.77E-05
F1	0.41	2.18E-05
F16	0.40	3.13E-05
F12	0.38	7.61E-05
F2	0.38	1.00E-04
F15	-0.33	6.65E-04
F7	-0.32	1.14E-03
F10	-0.31	1.41E-03
F17	-0.19	5.12E-02
F3	0.17	9.76E-02
F9	0.16	1.06E-01
F8	0.08	4.08E-01

Table 5.12: Correlation of individual feature measurements with Strain score.

The effectiveness of this feature selection procedure is likely to be affected by the following considerations:

1. There is likely to be some non-linearity in the relationship between feature measurements and GRBAS scores. Pearson Correlation is linear and may not accurately reflect this non-linearity. The Pearson correlation coefficient is a measure of the strength of a linear model of this relationship. An alternative measure may be more appropriate and will be investigated in further analysis as a wrapper method based on NRMSE.

2. Removing a feature may improve the score prediction made according to the remaining features. This can observed in Table 5.3 where, in Grade prediction, using 18 features gives a better prediction than using 20 features

5.4.1 Improvement in performance of the prediction models

The following five different approaches may be considered for improving the performance of the prediction models:

1. Reducing the complexity of the prediction models by reducing their dimensionality, on the grounds that extra dimensions may confuse the learning algorithms and may cause it to have high variance. Dimensionality reduction will also reduce the computational complexity of the models.

2. Improving the MLR or KNNR algorithms.

3. Improving the accuracy of the feature measurements.

4. Including more or different features such as average jitter and average shimmer.

5. Increasing the data-base size or introducing extra scorers.

Item 5 is not feasible for this project in view of the resources and time that would be involved. Item 2 to 4 are ongoing topics. Item 1 is the subject of the next section.

5.5 Dimensionality reduction

Dimensionality reduction is the process of reducing the number of features that need to taken into account when making predictions. Dimensionality reduction methods can be classified as ‘feature extraction’ or ‘feature selection’ methods. Feature extraction methods transform a high-dimensional space of feature measurements to a space of fewer dimensions. Principal Component Analysis (PCA) can achieve such a transformation without reference to the output scores by exploiting dependencies between feature measurements, that may be assumed to be linear. Many other transformations exist which can exploit nonlinear dependencies and the statistics of the output scores [Sam06, DHZS02]. Feature extraction has advantages, such as computational complexity reduction, which have not been exploited in this thesis.

‘Feature selection’ approaches try to find a subset of the original variables that enable more accurate prediction by the elimination of irrelevant and confusing information. Direct feature selection searches to identify, individually, the relevant features and discard the irrelevant ones. Feature selection may be carried out in combination with feature extraction as mentioned earlier. Such methods are instances of a wide range of general strategies for dimensionality reduction, which seek to map the input variables into a lower dimensional space prior to running the supervised learning algorithm. In this thesis, the main goal of direct feature selection is obtaining a subset of features that produces lowest error on the regression models. A learning algorithm is faced with the problem of selecting a relevant subset of features which makes the best prediction while ignoring the rest in the features. To achieve the best possible performance with MLR and KNNR on a particular training set, a feature subset selection method should consider how the algorithm and the training set interact. Since the usual goal of supervised learning algorithms is to minimise regression error on an unseen test set, we have adopted this as our goal in guiding the feature subset selection. There are three approaches; filter, wrappers and embedded methods [YTGF99]. Filter and wrappers are used in the experimental methodology in this thesis.

5.5.1 Filter methods

Filter methods are generally applied as pre-processing steps, with subset selection procedures that are independent of the learning algorithm. Filter methods apply ranking to the features. The ranking denotes how useful each feature is likely to be for the regression models. Once this ranking has been computed, a feature sub-set composed of the best features is created. Although this leads to a faster learning process, it is possible for the criterion used in the pre-processing step to result in a subset that may not work very well downstream in the learning algorithm. However, filter methods are fast to compute, and can be successful in capturing useful information in the feature set and eliminating the rest [YTGF99]. Filter method can be classified as being based on ‘distance or separability’ measures, or ‘correlation and information theoretic’ measures.

5.5.2 Feature ranking

One of the simplest ranking methods is to use the Pearson correlation coefficient between each feature and the reference GRBAS scores in data-base. Tables (5.8) to(5.12) show the 20 available features with their Pearson correlation coefficients with respect to ‘G’, ‘R’, ‘B’, ‘A’ and ‘S’ respectively. In each table, the features are re-ordered according to their values of correlation, and the P-values are shown. A P-Value less than 0.05 means that the probability of obtaining the correlation results by chance is low.

5.5.3 Feature reduction

Feature reduction may be carried out with respect to the Pearson correlation of each feature with a GRBAS component. As seen in Table (5.8), F8 has the lowest correlation with Grade, therefore this is ignored when 19 features are used for the prediction.

5.5.4 Wrapper methods

Wrapper methods train a new model for each possible combination of features. With such methods, the performances of different combinations of features for GRBAS prediction may be investigated [KJ97]. They are computationally intensive, but usually provide the best performing feature subset. In these methods, the subset selection takes place based as part of the learning algorithm (in our case MLR or KNNR) used to train the model itself. The subset that is selected decided in the context of the learning algorithm. The main idea of feature subset selection is to exclude redundant or irrelevant features that can lead to an increase of the regression error and increase the computational cost. ‘Greedy Search’ and ‘Exhaustive Search’ are two common wrapper methods. They are described in the following sections.

5.5.4.1 ‘Greedy search’ selection

There are two different types of ‘greedy search’ wrapper methods for feature selection: ‘forward selection’ and ‘backward elimination’. Backward elimination starts off using all feature variables and progressively eliminates the least promising ones. Forward selection progressively incorporates feature variables into larger and larger subsets. When used with training by cross-validation, at each validation stage, greedy search methods work by proposing changes to a current subset of features and accepting these changes only if they result in a reduction in the validation prediction error, i.e. the error averaged over the 'set aside' fold [GE03].

Greedy forward selection was investigated as a method of finding the best feature-subset as part of the model training procedure for GRBAS prediction. The procedure started with an empty subset and incrementally introduced features one by one. Each new feature was included on a 'probationary' basis and the resulting regression model was evaluated as now described.

The evaluation required the model to be trained for all stages of a cross-validation process. At each stage, all training subjects were used apart from those in the in the 'set aside' validation fold. At each stage, the NRMSE over the 'set-aside' validation fold was calculated and after completing all cross-validation stages an averaged validation NRMSE was calculated for the probationary feature set.

A probationary feature is kept only if there is a significant reduction in the validation NRMSE values averaged over all stages. Where several probationary features achieve a significant NRMSE reduction, the feature which achieves the greatest reduction is chosen. If all possible probationary features fail to achieve a significant reduction, the search terminates and the subset is considered complete. This process is continued by selecting and evaluation the effect of further probationary features until either a complete subset is found of the subset contains a maximum of ten features. This process produces a feature-subset with up to ten features.

To test the effectiveness of this approach, it is repeated for 20 trials with different random choices of 80 training subjects and 22 'set aside' testing subjects. Figure 5.6 shows the performed process for doing the 'Greedy Search'.

This 'greedy' approach does not consider all possible subsets and therefore may not find the optimum feature subset. However, by identifying which features cause a decrease in prediction error when augmenting certain subsets, it can make useful selections of feature subsets. Greedy search approaches have advantages over exhaustive search methods considered next in that they are computationally less intensive. When used with MLR model training, repeated MLR calculations have to be carried out using the pseudo-inverse Equation (5.7). When used with KNNR, the optimum K calculation procedure outlined in Section 5.2.1 must be carried out along-side the greedy search

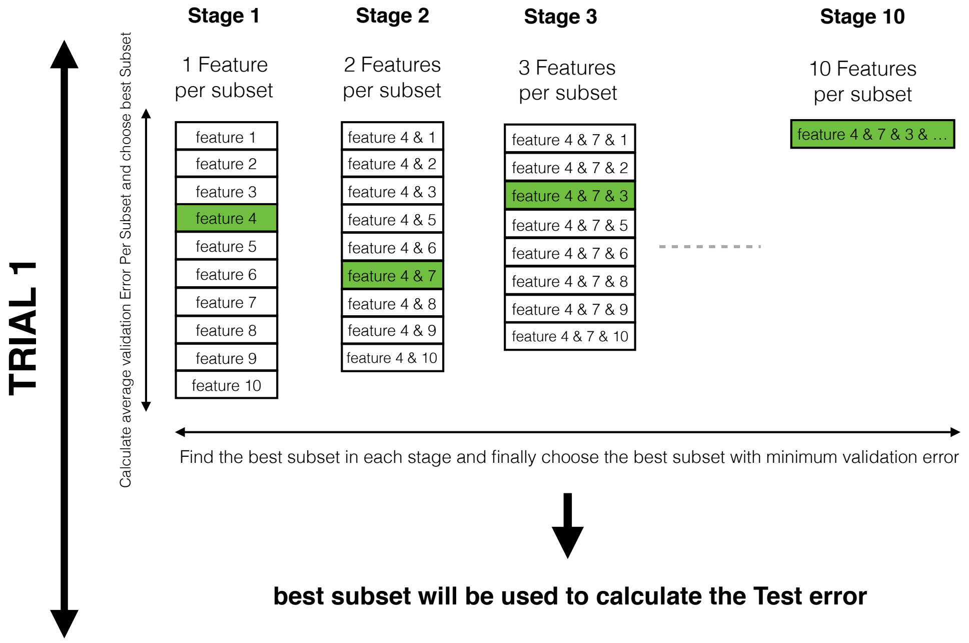


Figure 5.6: Greedy forward selection by 10-fold cross validation. The procedure started with an empty subset and incrementally introduced features one by one.

procedure.

5.5.4.2 'Exhaustive search' selection

Exhaustive search methods aim to consider every possible subset of features to find which one gives the best result prediction. This 'brute force' approach is possible with a small number of features, but can require massive amounts of computation for larger numbers of features. In this work, the computation time was made manageable by pre-selecting, for each GRBAS component, ten out of the 20 available features as described in Section 5.4. This was done by discarding the features with lowest Pearson correlation to the GRBAS scores to produce Tables 5.3 to 5.7. It could have been done using the 'backward elimination' approach to 'greedy selection' as described above, though with greater computational cost. Experiment results by [BS92] show the K-fold-cross-validation has better performance than leave-one-out-cross validation (LOOCV) for feature selection in linear regression. After the pre-selection, the cross-validation methodology described below may be followed to train the predictor by exhaustive search feature selection.

1. With ten pre-selected features there are $2^{10} - 1 = 1023$ different subsets of features. Each of these subsets must be used to train an MLR model and populate a KNNR model with a sub-set of training subjects chosen for cross-validation as described in Section 5.2. The sub-set of training subject omits one fold' set aside for validation. Once trained, the models must be 'validated' for subjects in the 'set aside' fold to obtain a value of NRMSE for each model. This procedure is repeated for each of the 'set aside' folds, for each model. The values of NRMSE for each model are then averaged to obtain a validation-averaged NRMSE for each model, for each feature-subset. For each model and for each GRBAS component, the 1023 feature-subset are used to train and then test a predictor for the'set-aside' testing subjects to obtain a testing-NRMSE for trial 1.

2. Steps 1 and 2 as outlined above are repeated for a second trial with a different random choice of training subjects and 'set aside' testing subjects. A testing-NRMSE for this second trial is thus obtained.

3. In total, sufficient trials must be carried out to obtain an average of trial-averaged-validation-NRMSE and an average of trial- testing-NRMSE for each model and for each GRBAS component. This gives a measure of the effectiveness of the training procedure which includes the selection of the best feature-subset.

4. The feature selection was made for the subset with the lowest the average of trial-averaged-validation-NRMSE.

5. The average of trial-testing-NRMSE was obtained for those selection of features.

6. Figure 5.7 shows the performed process for doing the 'exhaustive search'. In each trial 10-Fold cross validation was carried out.

5.5.4.3 Parallel computing in exhaustive search

Fortunately, the exhaustive search procedure is well suited to parallel computation because several independent searches can be carried out at the same time. To reduce computation time for the 'Exhaustive Search' method, parallel computing was performed in MATLAB.'Parfor' function was used in MATLAB to apply parallel computing. Normal computation for the 'exhaustive search' for each GRBAS component takes about 7-8 hours, doing 'parallel computing' with 8 CPU cores reduced the computation time to about 1-2 hours.

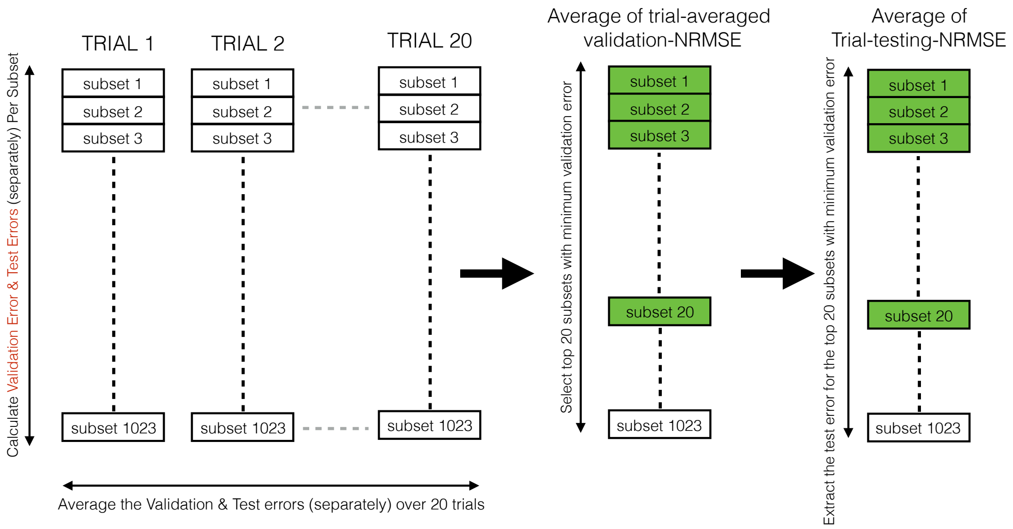

TRIAL 1
TRIAL 2
TRIAL 20
Average of trial-averaged validation-NRMSE
Average of Trial-testing-NRMSE
subset 1
subset 2
subset 3
subset 20
subset 1023
Calculate Validation Error & Test Errors (separately) Per Subset
Average the Validation & Test errors (separately) over 20 trials
Select top 20 subsets with minimum validation error
Extract the test error for the top 20 subsets with minimum validation error

Figure 5.7: Feature Selection by exhaustive search. In each trial 10-Fold cross validation was carried out for 1023 different subsets of features.

5.6 Grade Prediction

This section presents results obtained by applying 'greedy forward selection' and 'exhaustive search selection' for 'Grade' training and testing the models.

5.6.1 Greedy forward selection for 'Grade'

The greedy forward selection process that may be used to augment the training of MLR and KNNR regression models was explained in Section 5.5.4.1. For each model, for each GRBAS component, we first train the model using each of the 10 features in turn as a single feature. The model is then evaluated for each choice of feature over all ten cross-validation stages to obtain ten values of validation-NRMSE, which are averaged. The best single feature in terms of the lowest averaged validation error is thus identified and constitutes the starting subset for each model. We now try to increase the feature subset size by adding another feature. This is done by introducing each of the remaining features in turn on a probationary basis. The feature which gives the largest significant reduction in the averaged validation-NRMSE becomes a permanent member of the feature subset. If a significant reduction is not achieved by adding any of the remaining features to the subset, the subset is considered complete. If a significant reduction is achieved, the process continues to try to find a third feature for the subset, and so on up to a maximum subset size of ten features. This cross-validation procedure produces a feature-subset with up to ten features.

To test the effectiveness of this approach, it is repeated for 20 trials, each with a different random choice of 80 training subjects and 22 ‘set aside’ testing subjects. For each trial, the ‘greedy forward selection’ cross-validation process described above is carried out to find a suitable feature subset, and then the model with this feature subset is tested for the 22 ‘set aside’ subjects.

Figure 5.8 and 5.9 depict the trial-testing-NRMSE obtained for the best feature-subset for each of the 20 trials, for MLR and KNNR respectively.

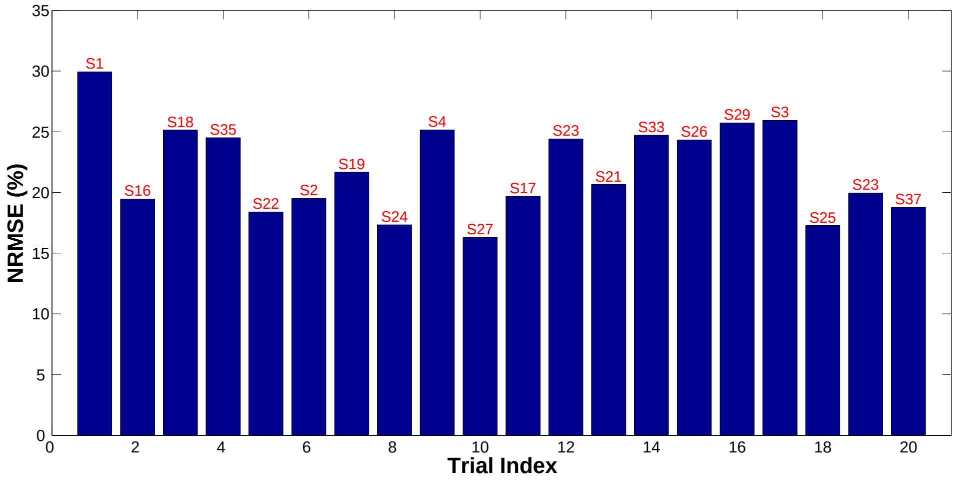


Figure 5.8: Trial-Testing-NRMSE for the best selected feature subset (MLR-Grade-Greedy). S labels are features subsets in Table E.1

Table E.1 in the Appendix defines the feature subsets S1, S2, etc. referred to in these graphs. Each trial uses the best feature-subset found (by greedy forward selection) for that trial, and it may be different for each trial. The success of the training method (which includes the method of choosing a feature-subset) may be judged from the results of the 20 trials.

It may be seen in 5.8 and 5.9 that for each trial, a feature sub-set has been selected, giving a NRMSE value around 21% which means an average grade score error of about 0.63 (over the scale 0 to 3). It is interesting that the performances of MLR and KNNR are so close. If this indication of the accuracy of greedy forward selection is considered acceptable, or the best we can get, the final choice of feature-subset is made by repeating the ‘greedy forward selection’ training process but this time using all 102 subjects for ‘cross-validation’ training. We do this to try to get the best possible result from the method we have just tested.

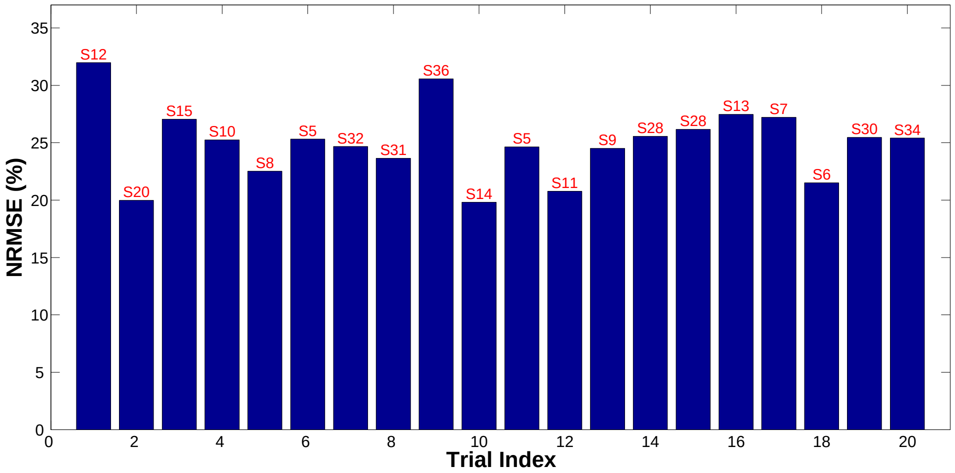


Figure 5.9: Trial-Testing-NRMSE for the best selected feature subset (KNNR-Grade-Greedy). S labels are features subsets in Table E.1

The final result was not tested as we have no further subjects. But the choice has been made by a method that has been validated and tested so we may have confidence in its suitability. Greedy forward selection is not the best possible method for feature-subset selection because it does not examine all possibilities. The test results in Figures 5.8 and 5.9 give the best obtainable feature subset for each trial, but they do not tell us how the best one for each trial performs in the other trials. This omission could be remedied by recording how all subsets perform in each trial. Also, the final system could be tested with the optimal final choice of feature-set (now fixed) in a further series of 20 trials. However this would be slightly suspect as the test data would include that used to derive the best feature sub-set. Instead of doing this, we preferred to concentrate on the exhaustive search method which promises better results.

5.6.2 Exhaustive search for ‘Grade’

The Exhaustive Search procedure was explained in Section 5.5.4.2. After discarding the 10 features with lowest Pearson correlation with Grade as described in Section 5.4, the rest of the features were pre-selected for an ‘exhaustive search’. The ten pre-selected features are the first ten as presented in Table (5.8), i.e. F4, F14, F20, F11, F13, F1, F5, F2, F19 and F6. With ten features there are $1{+}10 + 10 \times 9/2! + \ldots +10{=}$ $2^{10} - 1 = 1023$ different subsets of features (excluding the empty subset). For each subset, the MLR and KNNR models were trained using (or populated by) 72 subjects in nine cross-validation folds and evaluated (validated) by the tenth fold, this process being repeated for all ten permutations of folds. The validation-NRMSE was then computed between the predicted Grade scores and the corresponding reference scores for the cross-validation samples and averaged across the ten cross-validation steps to obtain an validation-averaged-NRMSE. Then all 1023 subset tested using the 'set-aside' testing sample of 22 subjects; this procedure being repeated for 20 trials. Then the average of trial- averaged-validation-NRMSE was obtained for all 1023 subset. A selection of feature subset was made, and then then averaged of trial-tesing-NRMSE was obtained for the selected feature subset.

For 'Grade', box-plots for the 20 best performing feature-subsets, as selected from the 1023 possible feature-subsets, are presented in 5.10 (for MLR) and 5.11 (for KNNR). The 'top twenty' best performing subsets are those with the lowest average of trial-averaged-validation-NRMSE. The labels S1, S2, etc. refer to feature subsets as defined in Table E.2 in the appendix. The range of average of trial testing-NRMSE values for the best twenty subsets are indicated. The box-plots show how each 'top-20' subset behaves over the whole trial.

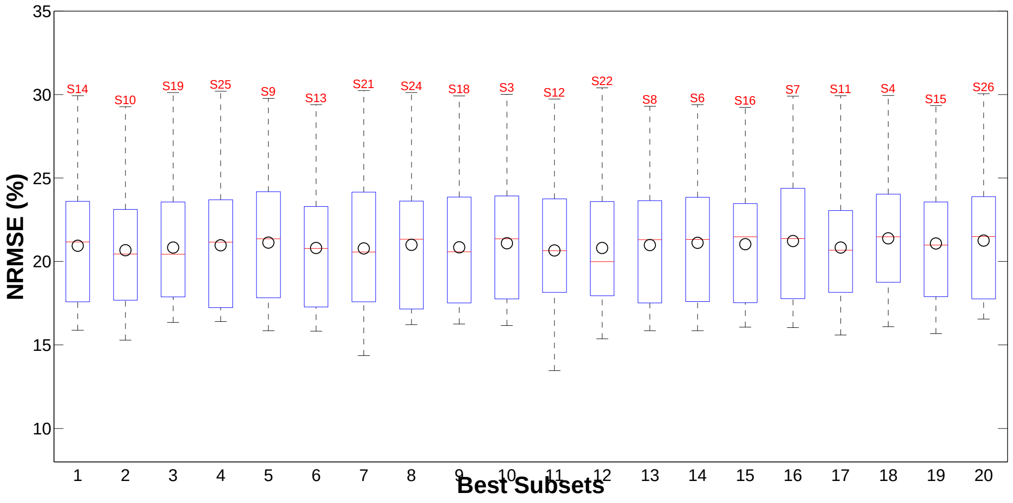


Figure 5.10: Average of Trial-Testing-NRMSE for the best 20 selected feature (MLR-Grade-EXHAUSTIVE). S labels are feature subsets in Table E.2. ('o':average, '-':median, '+':outliers)

'Box plots' are useful for identifying outliers as well as for comparing distributions of average of trial-testing-NRMSE values. A median, maximum and minimum for each of these 'top 20' subsets is indicated. It is interesting that the variation in the median, and even the range is quite small for MLR, and a little larger for KNNR.

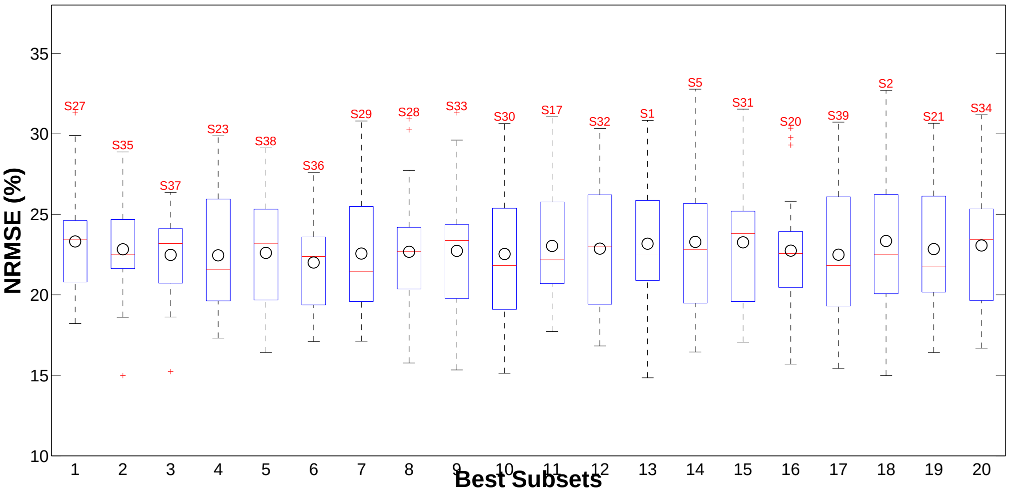


Figure 5.11: Average of Trial-Testing-NRMSE for the best 20 selected feature subset (KNNR-Grade-EXHAUSTIVE). S labels are feature subsets in Table E.2. (‘o’:average, ‘-’:median, ‘+’:outliers)

The lowest value of average of trial testing NRMSE is 20% as obtained for feature-set S14 (for MLR) and 23% with feature-set S27 for KNNR. The ‘o’ in the boxes indicates the average of trial-testing-NRMSE for the ‘top 20’ subsets. Each subset has a different average, median, max and min NRMSE over 20 trials. The feature subset with the lowest average of trial-averaged testing-NRMSE is feature-subset ‘S14’ for MLR and ‘S27’ for KNNR . According to Table E.2, these subsets contain six features, i.e. F4, F14, F20, F1, F5 and F19 (MLR) and five features, i.e. F4, F14, F1, F5, F2 and F6 (KNNR). MLR has a lower minimum-NRMSE than KNNR. The ‘+’ show the obtained outlier in the NRMSE. Outliers are observed in subset S20, S27, S28, S33, S35, and S37 for (KNNR). These outliers increase error variance and reduce the power of statistical tests. If NRMSE is non-randomly distributed they can decrease normality. They can seriously bias or influence estimates that may be of substantive interest [Ras88, SM82, Zim94]. The linear and non-linear relation between the features may cause the detection of these outliers. As well as doing this, all 1023 feature-subsets were tested in 20 testing trials to obtain Figures (5.12) (for MLR) and (5.13).

Each figure shows the value of average of trial-testing-NRMSE that was obtained for each of the 1023 different feature-subsets. The indices of these subsets were re-ordered to achieve an increasing magnitude of the average of trial NRMSE. The red circle shows the obtained average NRMSE over 20 trials when all ten pre-selected features were used for the prediction and the green circle depict the testing error for

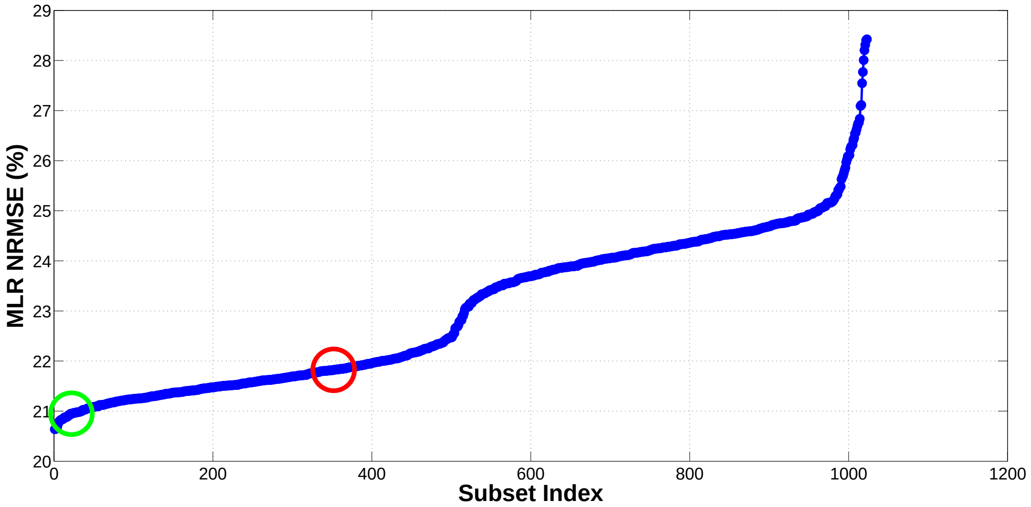


Figure 5.12: Average of Trial-Testing-NRMSE for all subset of features (MLR-Grade-EXHAUSTIVE). (green circle : best subset and red circle: 10 features)

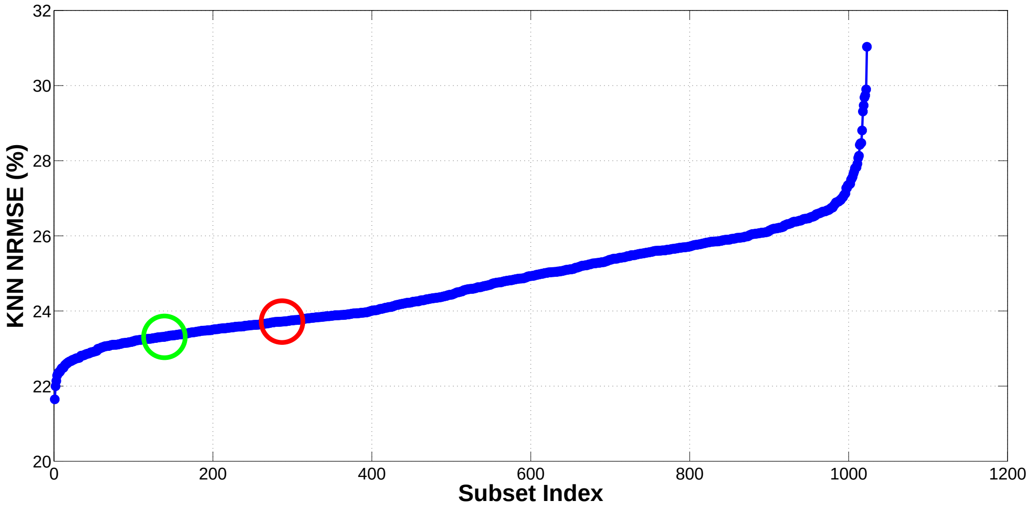


Figure 5.13: Average of Trial-Testing-NRMSE for all subset of features (KNNR-Grade-EXHAUSTIVE). (green circle : best subset and red circle: 10 features)

the best selected subset. These curves demonstrate that improvement and deterioration in the prediction accuracy of either method, compared with what is achieved using all ten features, can be achieved by feature selection. While significant deterioration can occur, it seems that the scope for improvement is rather small, i.e. about 2%.

5.6.3 Optimum K for Grade prediction by KNNR

As mentioned earlier, with KNNR, the NRMSE of the regression will be affected by the feature subset and value of K, which is the number of nearest neighbours chosen. For each subset the validation-NRMSE was averaged over 10 folds for K, 1 to 10 in Trial 1. The tesing-NRMSE was obtained for each subset for K, 1 to 10. This is repeated for a second trial with a different random choice of training subjects and set aside testing subjects. A testing-NRMSE for this second trial is thus obtained. This process was repeated over 20 trials.

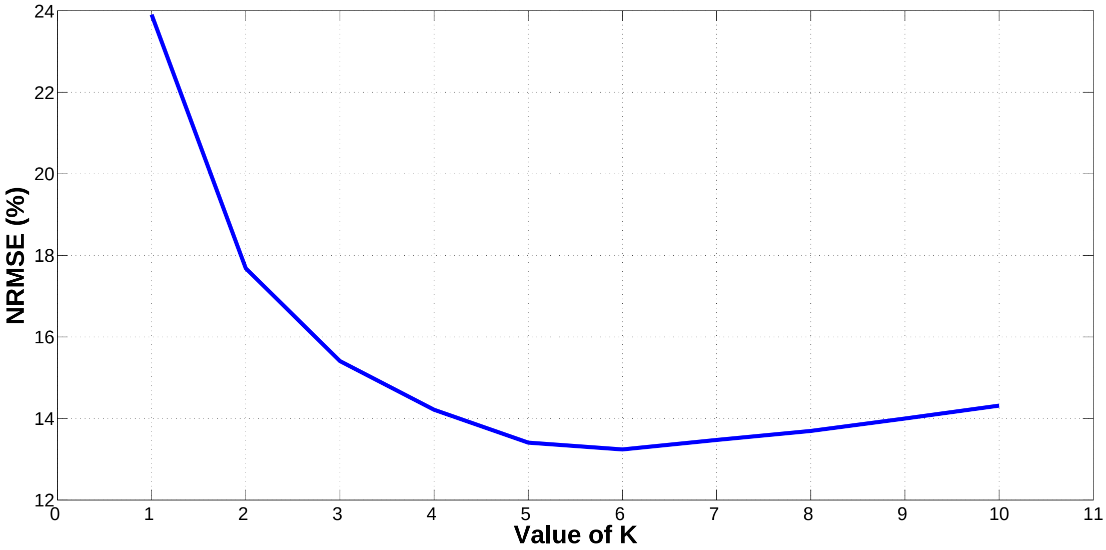


Figure 5.14: Average of Trial-Averaged-Val-NRMSE Per K Best Subset (Grade)

The average of trial-averaged-validation-NRMSE and average of trial-testing-NRMSE were obtained for each subset of features for K, 1 to 10. A grid search [BB12] was used to find out the best feature subset and K with the lowest average of trial-averaged-validation-NRMSE amongst 1023 different subsets. K and the choice feature subset are therefore jointly selected to find the average of trial-testing-NRMSE.

Figure 5.14 shows the average of trial-averaged-validation-NRMSE for K 1 to 10 when the best subset (S27) was used for ‘Grade’ prediction. The best subset with the lowest average of trial-averaged-validation-NRMSE has K=6.

5.7 Roughness prediction

5.7.1 Greedy forward selection for 'Roughness'

The Greedy Forward Selection was described in Sections 5.5.4.1 and 5.6.1. The same procedure was used for 'Roughness' prediction.

Figures 5.15, 5.16 depict the trial- testing-NRMSE error obtained for MLR and KNNR. Table E.3 in the Appendix defines the feature subsets S1, S2, etc. referred to in these graphs. Each trial uses the best feature-subset found for that trial, and it may be different for each trial.

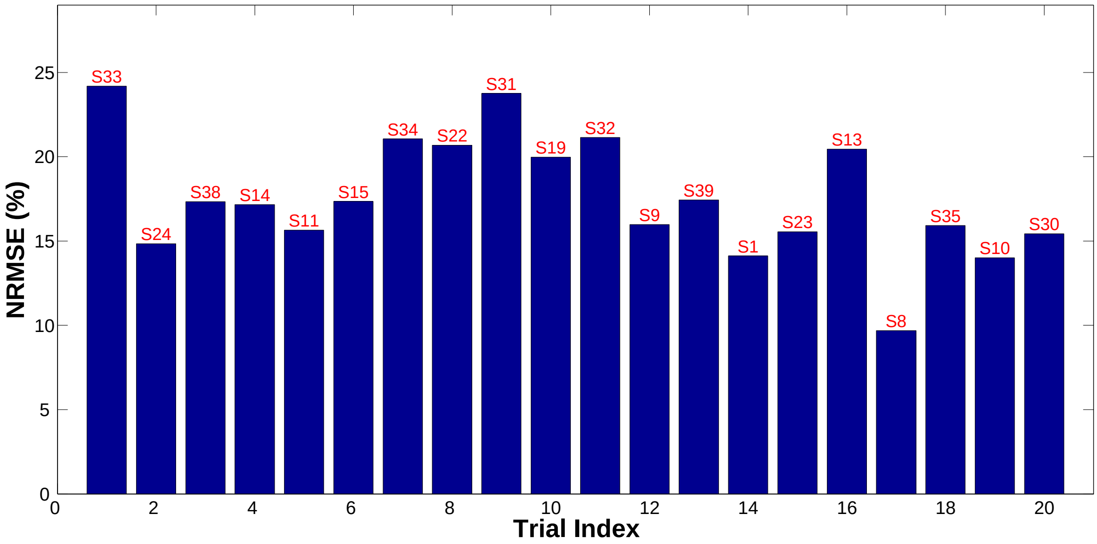


Figure 5.15: Trial-Testing-NRMSE for the best selected feature subset (MLR-Roughness-Greedy). S labels are features subsets in Table E.3

5.7.2 Exhaustive search for 'Roughness'

The Exhaustive Search procedure was explained in Section 5.5.4.2 and 5.6.2. After discarding the 10 features with lowest Pearson correlation with Roughness as described in Section 5.4 the rest of the features were pre-selected for an 'exhaustive search'. The ten selected features are the first ten as presented in Table (5.9), i.e. F4, F14, F20, F12, F11, F19, F5, F1, F7 and F2.

Figures 5.17 (for MLR) and 5.18 (for KNNR with optimum K) show the average of trial-testing-NRMSE that was obtained for 'top twenty'. The 'top twenty'

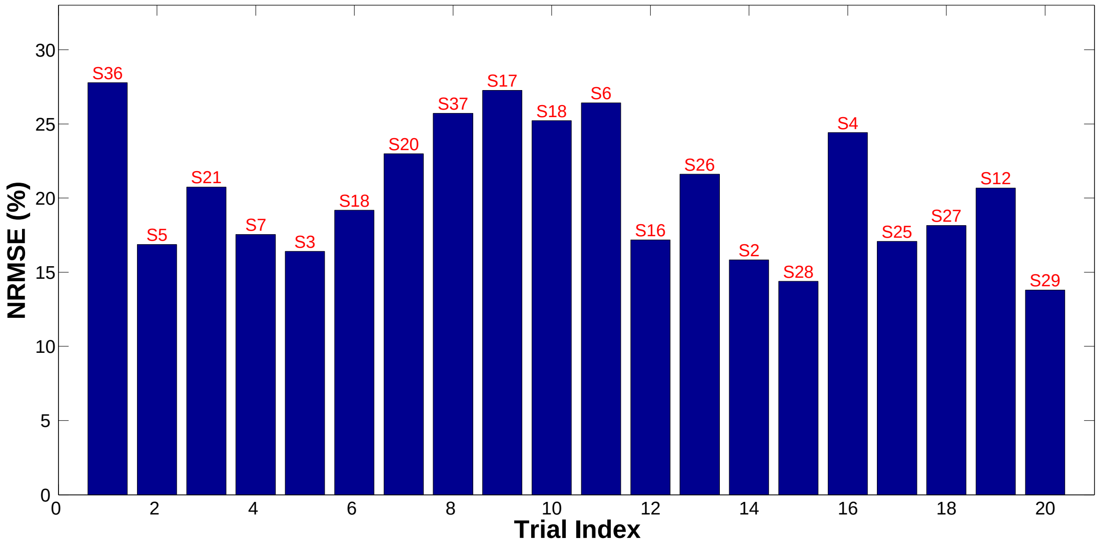


Figure 5.16: Trial-Testing-NRMSE for the best selected feature subset (KNNR-Roughness-Greedy). S labels are features subsets in Table E.3

best performing subsets are those with the lowest average of trial-averaged-validation-NRMSE. The labels S1, S2, etc. refer to feature subsets as defined in Table E.4 in the appendix.

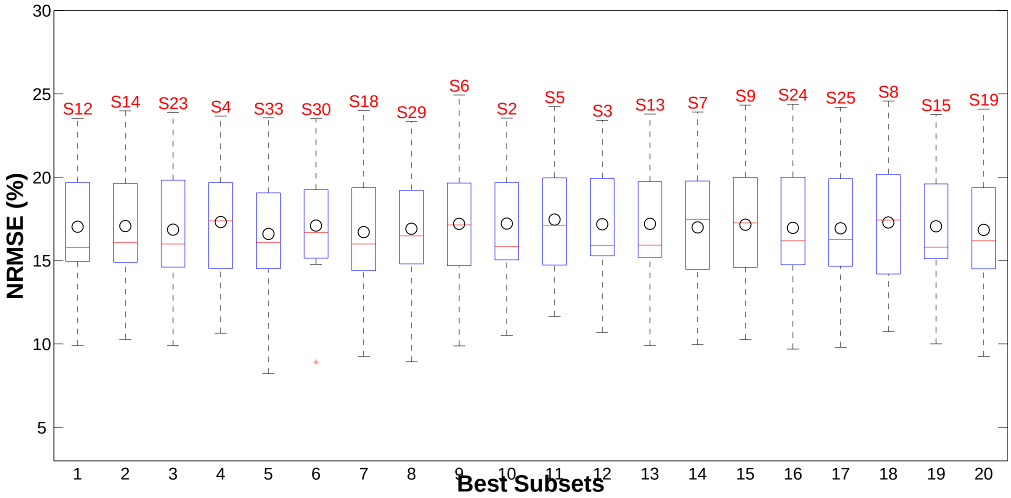


Figure 5.17: Average of Trial-Testing-NRMSE for the best 20 selected feature subset (MLR-Roughness-EXHAUSTIVE). S labels are feature subsets in Table E.4. (‘o’:average, ‘-’:median, ‘+’:outliers)

The lowest value of average of trial-testing NRMSE is 17% as obtained for feature-set S12 (for MLR) and 18% with feature-set S38 for KNNR. According to Table E.2,

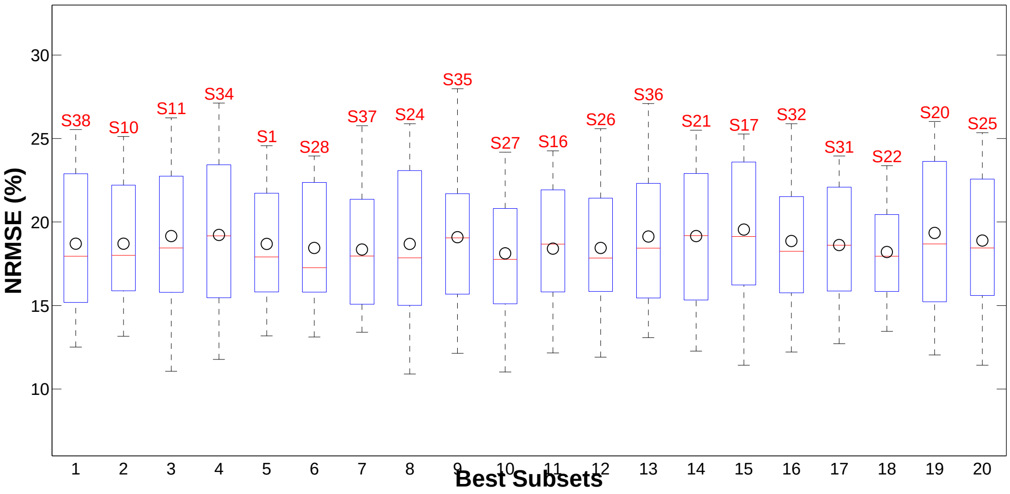


Figure 5.18: Average of Trial-Testing-NRMSE for the best 20 selected feature subset (KNNR-Roughness-EXHAUSTIVE). S labels are feature subsets in Table E.4. (‘o’:average, ‘-’:median, ‘+’:outliers)

these subsets contains five F4, F12, F5, F1 and F7 (MLR) and two features F4, F7 (KNNR). MLR has a lower minimum-NRMSE than KNNR. An outlier was observed in subset S30 (MLR). As well as doing this, all 1023 feature-subsets were tested in 20 testing trials to obtain Figures (5.19) (for MLR) and (5.20).

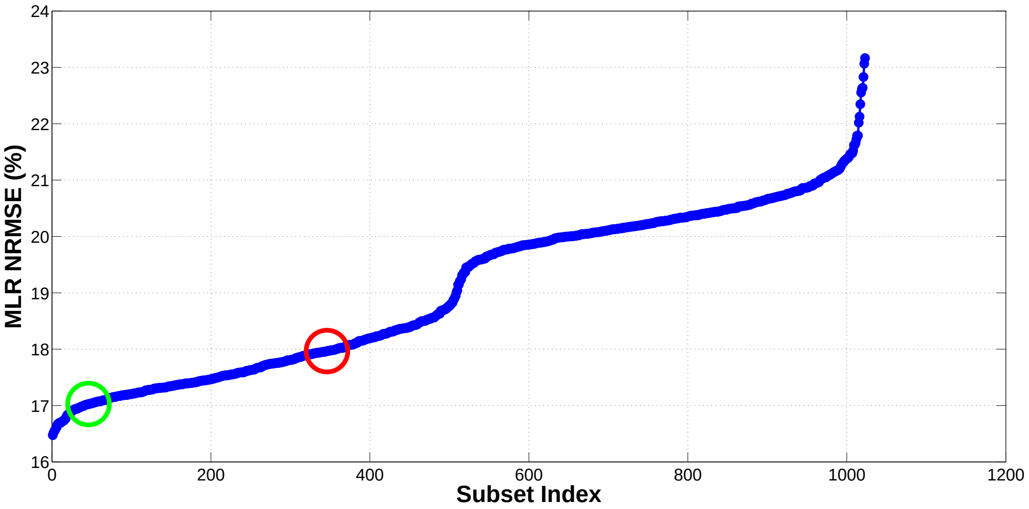


Figure 5.19: Average of Trial-Testing-NRMSE for all subset of features (MLR-Roughness-EXHAUSTIVE). (green circle: best subset and red circle: 10 features)

Each figure shows the value of average of trial testing-NRMSE that was obtained

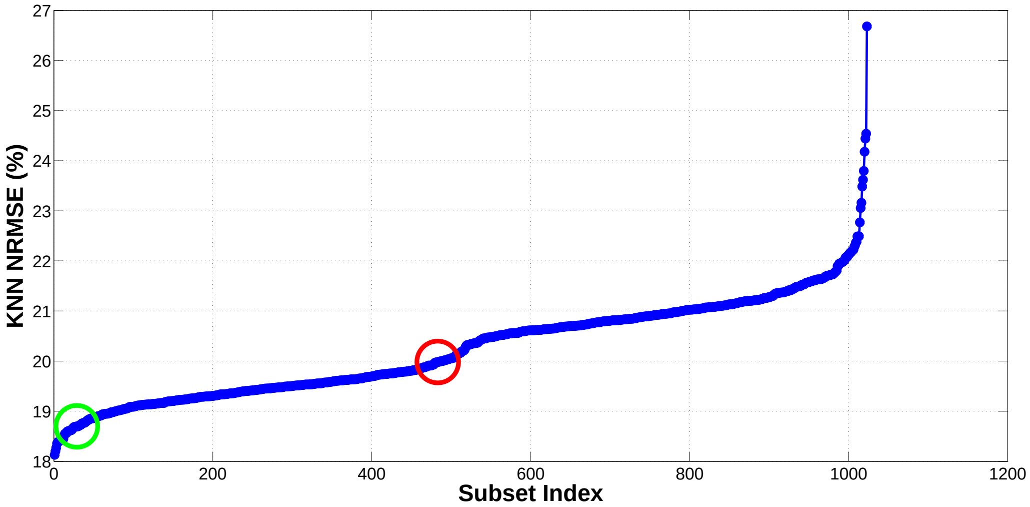


Figure 5.20: Average of Trial-Testing-NRMSE for all subset of features (KNNR-Roughness-Exhaustive). (green circle: best subset and red circle: 10 features)

for each of the 1023 different feature-subsets. These curves demonstrate that improvement and deterioration in the prediction accuracy of either method, compared with what is achieved using all ten features, can be achieved by feature selection. While significant deterioration can occur, it seems that the scope for improvement is rather small, i.e. of the order of about 2%.

5.7.3 Optimum K for Roughness prediction by KNNR

The K selection procedure was described in Section 5.6.3. Figure 5.21 shows the average of trial-averaged-validation-NRMSE in each trial for K, 1 to 10 when the best subset (S38) was used for 'Roughness' prediction. The best subset with the lowest average of trial-averaged-validation-NRMSE has K=10.

5.8 Breathiness prediction

5.8.1 Greedy forward selection for 'Breathiness'

The Greedy Forward Selection was described in Sections 5.5.4.1 and 5.6.1. The same procedure is used for 'Breathiness' prediction. Figures 5.22, 5.23 depict the average of trial-testing-NRMSE obtained for the best selected feature subset using MLR and KNNR.

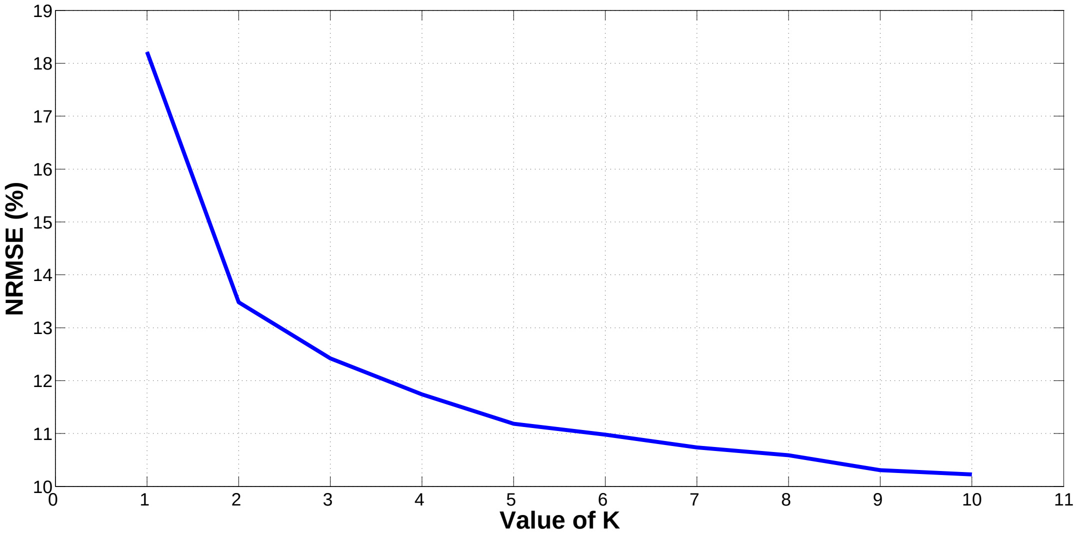


Figure 5.21: Average of Trial-Averaged-Val-NRMSE Per K for Best Subset (Roughness)

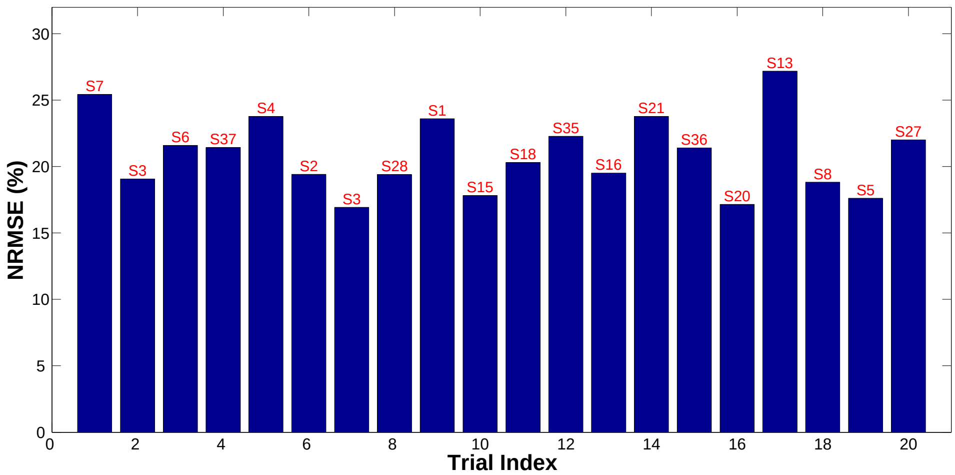


Figure 5.22: Trial-Testing_NRMSE for the best selected feature subset (MLR-Breathiness-Greedy). S labels are features subsets in Table E.5

Table E.5 in the Appendix defines the feature subsets S1, S2, etc. referred to in these graphs. Each trial uses the best feature-subset found for that trial, and it may be different for each trial.

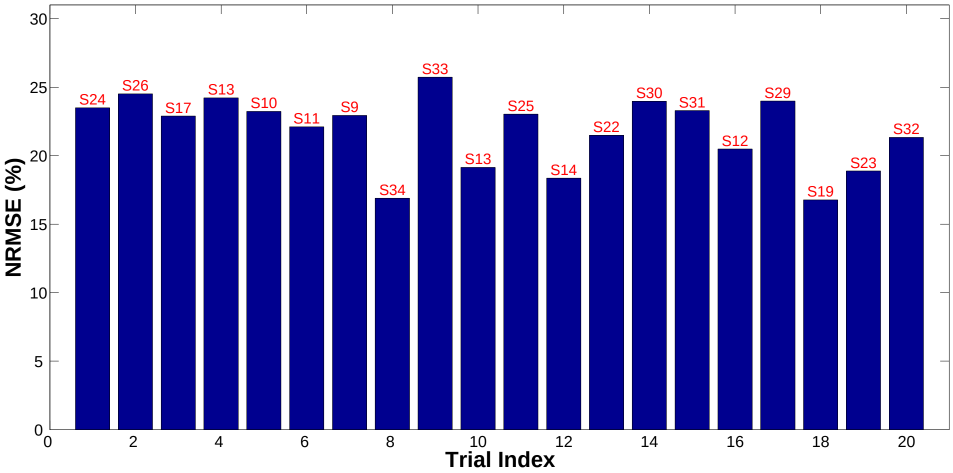


Figure 5.23: Trial-Testing_NRMSE for the best selected feature subset (KNNR-Breathiness-Greedy). S labels are features subsets in Table E.5

5.8.2 Exhaustive search for 'Breathiness'

The Exhaustive Search procedure was explained in Section 5.5.4.2 and 5.6.2. After discarding the 10 features with lowest Pearson correlation with Breathiness as described in Section 5.4 the rest of the features were pre-selected for an 'exhaustive search'. The ten selected features are the first ten as presented in Table (5.10), i.e. F13, F11, F4, F20, F14, F5, F2, F1, F7 and F10. Box-plots for the 20 best performing feature-subsets, as selected from the 1023 possible feature-subsets, are presented in Figure 5.24 (for MLR) and 5.25 (for KNNR).

The labels S1, S2, etc. refer to feature subsets as defined in Table E.6 in the appendix. The 'top twenty' best performing subsets are those with the lowest average of trial-averaged-validation-NRMSE. Figure 5.26 and 5.27 show the value of average of trial-testing-NRMSE that was obtained for each of the 1023 different feature-subsets.

These curves demonstrate that improvement and deterioration in the prediction accuracy of either method, compared with what is achieved using all ten features, can be achieved by feature selection. While significant deterioration can occur, it seems that the scope for improvement is rather small, i.e. of the order of about 1%.

5.8.3 Optimum K for Breathiness prediction by KNNR

The K selection for 'Breathiness' prediction was described in Section 5.6.3. Figure 5.28 shows the average of trial-averaged validation-NRMSE in each trial for K, 1 to

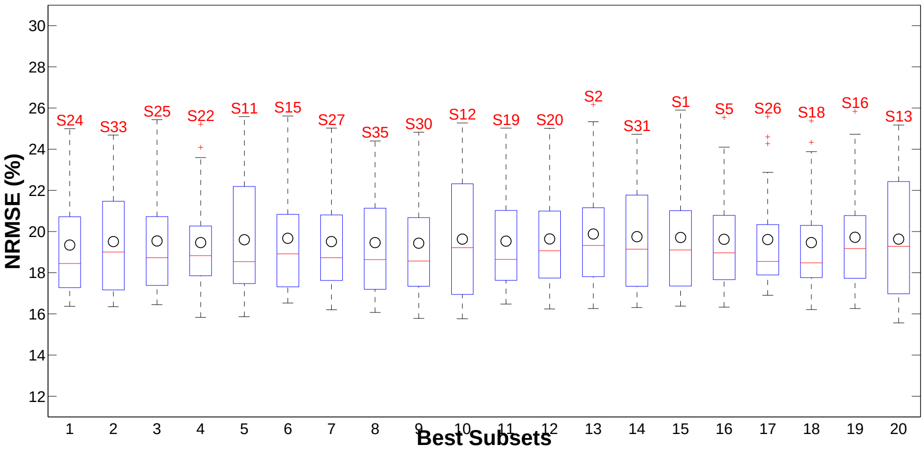


Figure 5.24: Average of Trial-Testing-NRMSE for the best 20 selected feature subset in each trial (MLR-Breathniess-EXHAUSTIVE). S labels are feature subsets in Table E.6. ('o':average, '-':median, '+':outliers)

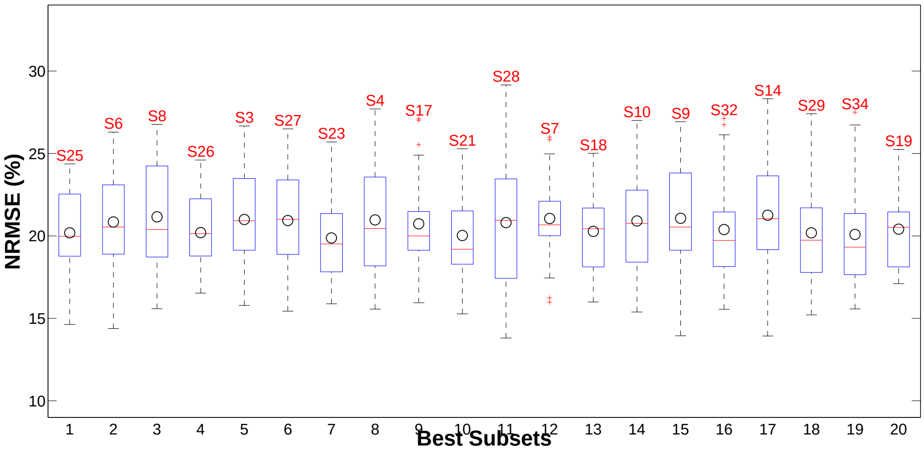


Figure 5.25: Average of Trial-Testing-NRMSE for the best 20 selected feature subset in each trial (KNNR-Breathniess-EXHAUSTIVE). S labels are feature subsets inTable E.6. ('o':average, '-':median, '+':outliers)

10 when the best subset (S25) was used for 'Breathiness' prediction. The best subset with the lowest average of trial-averaged- validation-NRMSE has K=5.

Average of Trial-Averaged-Val-NRMSE Per K Best Subset (Grade)

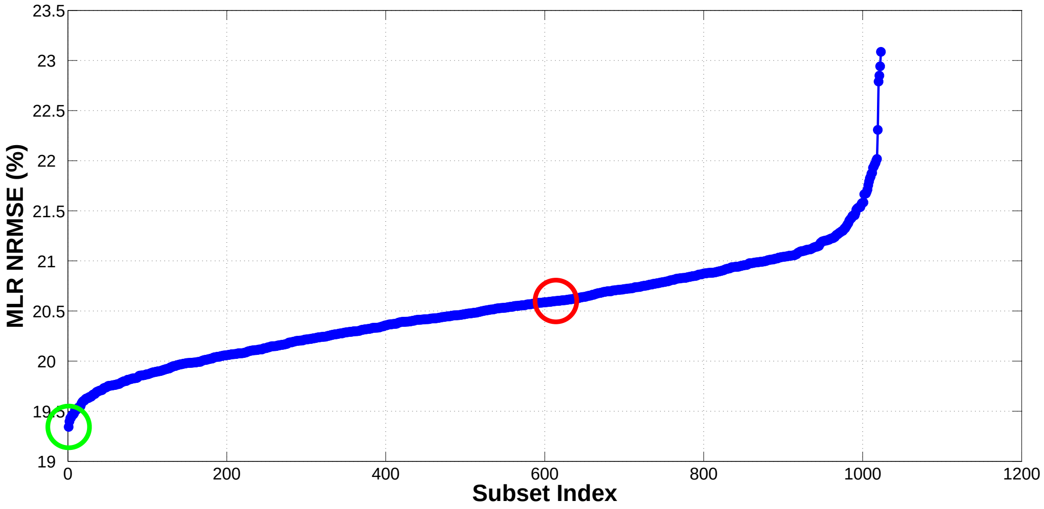


Figure 5.26: Average of Trial-Testing-NRMSE for all subset of features (MLR-Breathiness-EXHAUSTIVE). (green circle : best subset and red circle: 10 features)

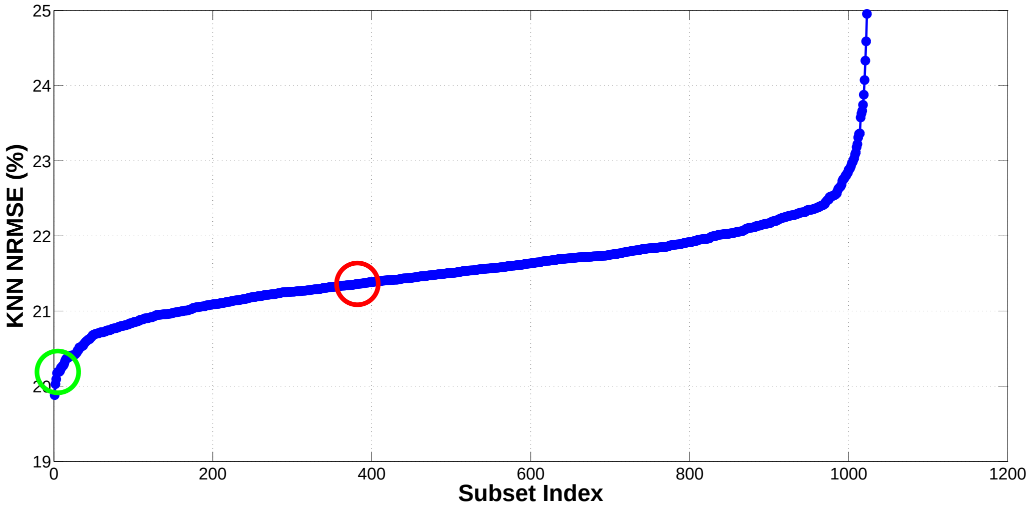


Figure 5.27: Average of Trial-Testing-NRMSE for all subset of features (KNNR-Breathiness-EXHAUSTIVE). (green circle : best subset and red circle: 10 features)

5.9 Asthenia Prediction

5.9.1 Greedy forward selection for ‘Asthenia’

The Greedy Forward Selection was described in Sections 5.5.4.1 and 5.6.1. The same procedure is applied for ‘Asthenia’ prediction.

Figures 5.29 and 5.30 depict the trial-testing-NRMSE obtained for MLR and KNNR. Table E.7 in the Appendix defines the feature subsets S1, S2, etc. referred to in these

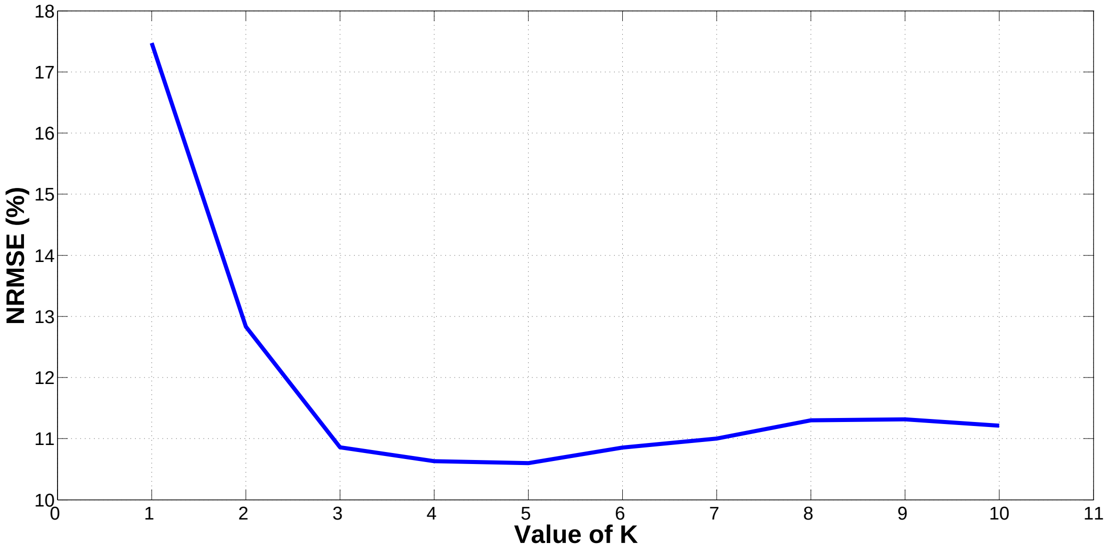


Figure 5.28: Average of Trial-Averaged-Val-NRMSE Per K for Best Subset (Breathiness)

graphs. Each trial uses the best feature-subset found for that trial, and it may be different for each trial.

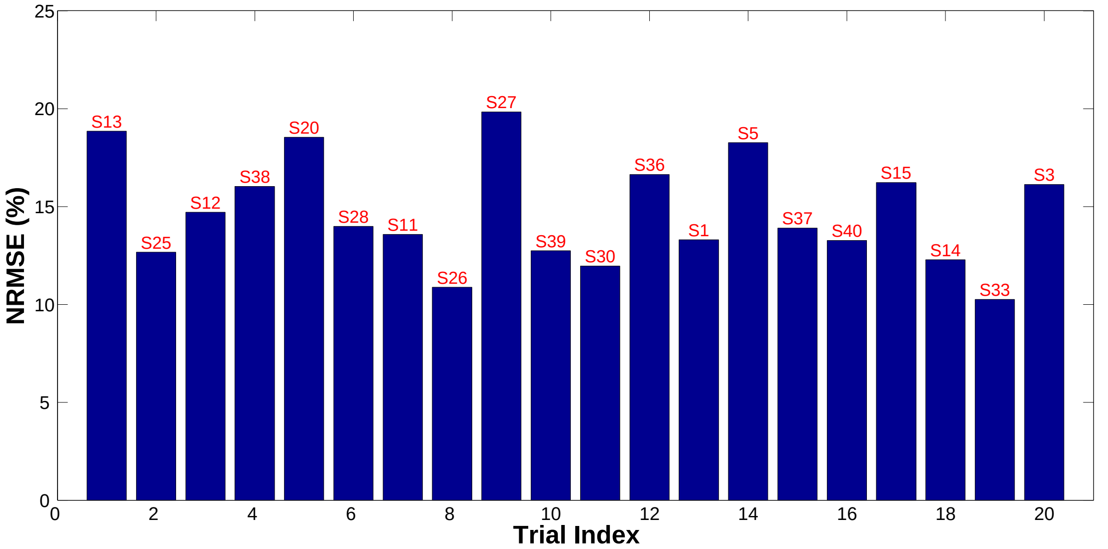


Figure 5.29: Trial-Testing-NRMSE for the best selected feature subset (MLR-Asthenia-Greedy). S labels are features subsets in Table E.7

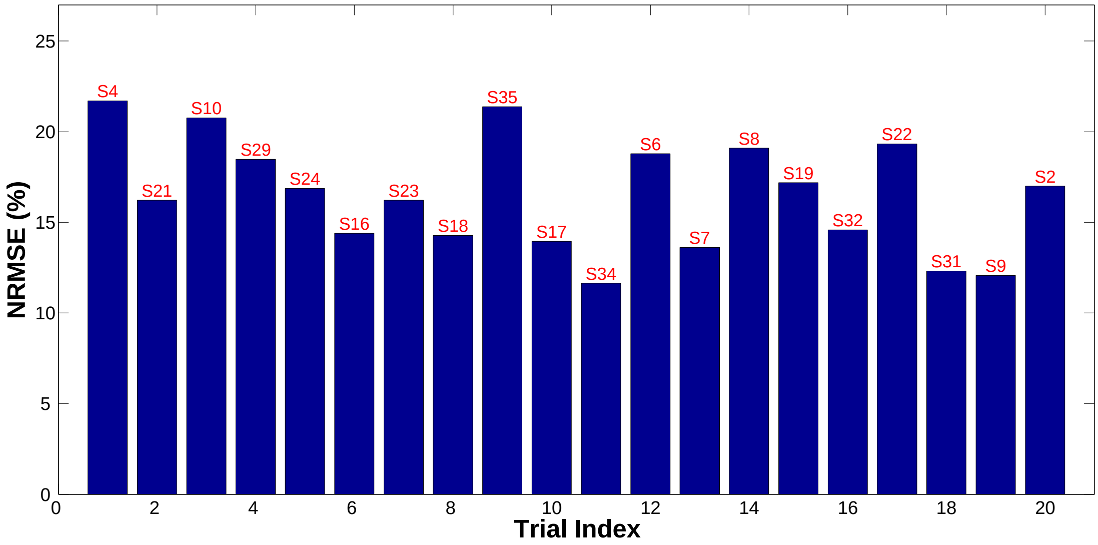


Figure 5.30: Trial-Testing-NRMSE for the best selected feature subset (KNNR-Asthenia-Greedy). S labels are features subsets in Table E.7

5.9.2 Exhaustive search for 'Asthenia'

The Exhaustive Search procedure was explained in Section 5.5.4.2 and 5.6.2.

After discarding the 10 features with lowest Pearson correlation with Asthenia as described in Section 5.4 the rest of the features were pre-selected for an 'exhaustive search'. The ten selected features are the first ten as presented in Table 5.11, i.e. F4, F11, F20, F14, F13, F19, F5, F16, F18 and F1.

Box-plots for the 20 best performing feature-subsets, as selected from the 1023 possible feature-subsets, are presented in Figure 5.31 (for MLR) and 5.32 (for KNNR).

The labels S1, S2, etc. refer to feature subsets as defined in Table E.8 in the appendix. The 'top twenty' best performing subsets are those with the lowest average of trial-averaged-validation-NRMSE. The feature subset with the lowest average of trial-averaged-validation-NRMSE is feature-subset 'S21' for MLR and 'S8' for KNNR.

According to Table E.8, these subsets contains seven F4, F11, F14, F13, F5, F18 and F1 (MLR) and five F4, F14, F13, F16, F1 (KNNR) features. Figures 5.33 (for MLR) and 5.34 (for KNNR with optimum K) show the value of average of trial-testing-NRMSE that was obtained for each of the 1023 different feature-subsets.

These curves demonstrate that improvement and deterioration in the prediction accuracy of either method, compared with what is achieved using all ten features, can be achieved by feature selection. While significant deterioration can occur, it seems that the scope for improvement is rather small, i.e. of the order of about 0.5%. The

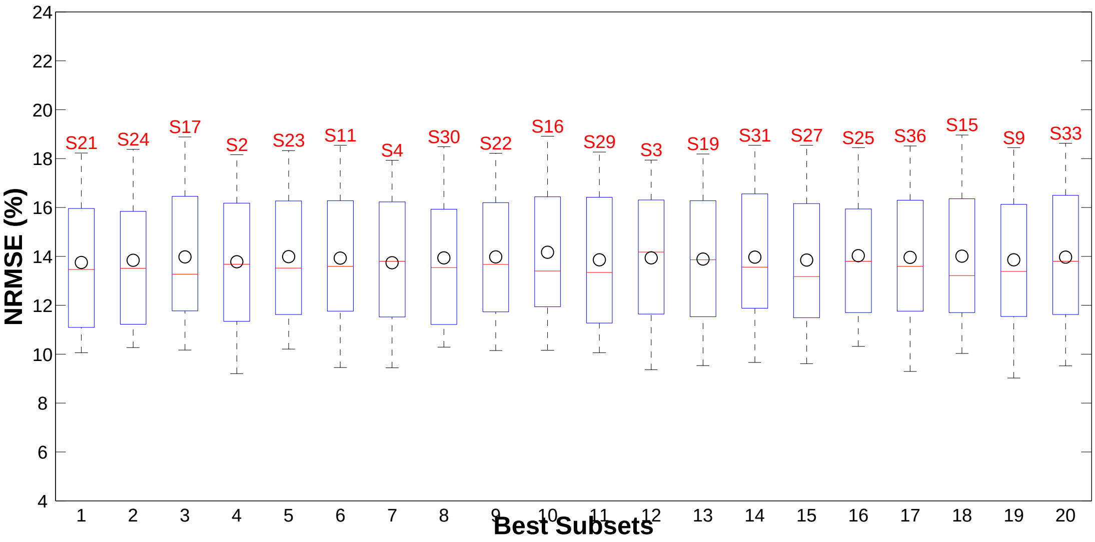


Figure 5.31: Average of Trial-Tesing-NRMSE for the best 20 selected feature subset (MLR-Asthenia-EXHAUSTIVE). S labels are feature subsets in Table E.8. (‘o’:average, ‘-’:median, ‘+’:outliers)

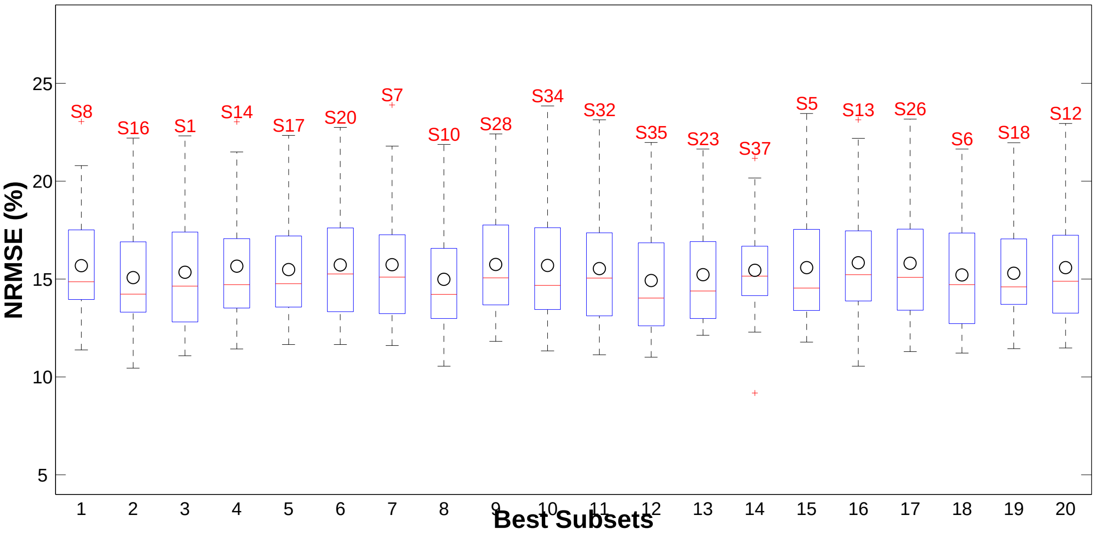


Figure 5.32: Average of Trial-Tesing-NRMSE for the best 20 selected feature subset (KNNR-Asthenia-EXHAUSTIVE). S labels are feature subsets in Table E.8. (‘o’:average, ‘-’:median, ‘+’:outliers)

observed outliers are subsets S7, S8, S13, S14 and S37 (KNNR).

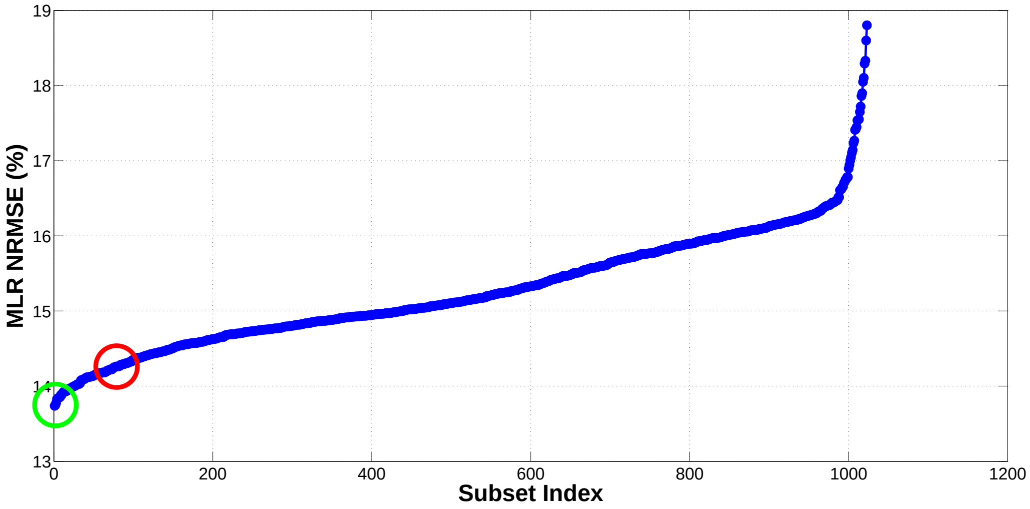


Figure 5.33: Average-Trial-Testing-NRMSE for all subset of features (MLR-Asthenia-EXHAUSTIVE). (green circle: best subset and red circle: 10 features)

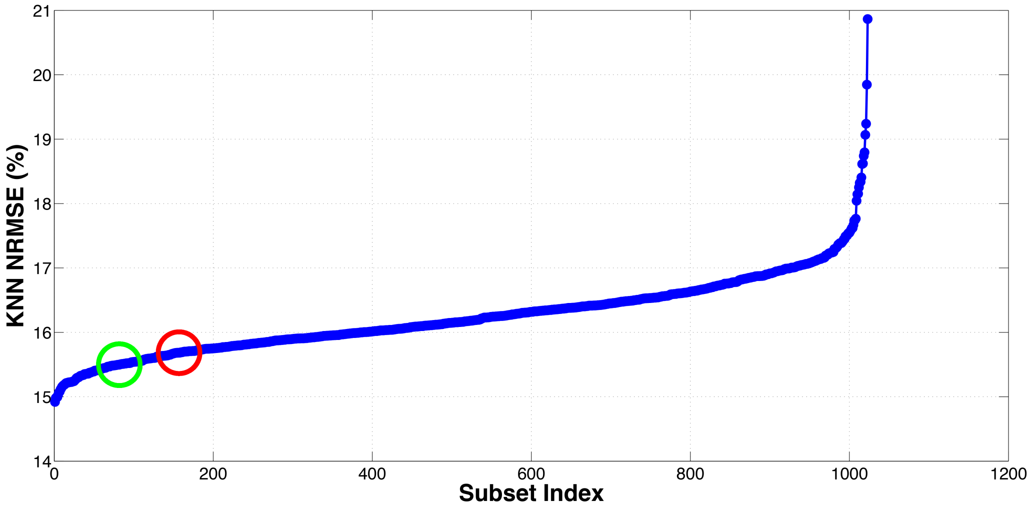


Figure 5.34: Average of Trial-Testing-NRMSE for all subset of features (KNNR-Asthenia-EXHAUSTIVE). (green circle : best subset and red circle: 10 features)

5.9.3 Optimum K for Asthenia prediction by KNNR

The K selection procedure was described in Section 5.6.3. Figure 5.35 shows the average of trial-averaged-validation-NRMSE in each trial for K, 1 to 10 when the best subset (S8) was used for 'Asthenia' prediction. The best subset with the lowest average of trial-averaged-validation-NRMSE has K=8.

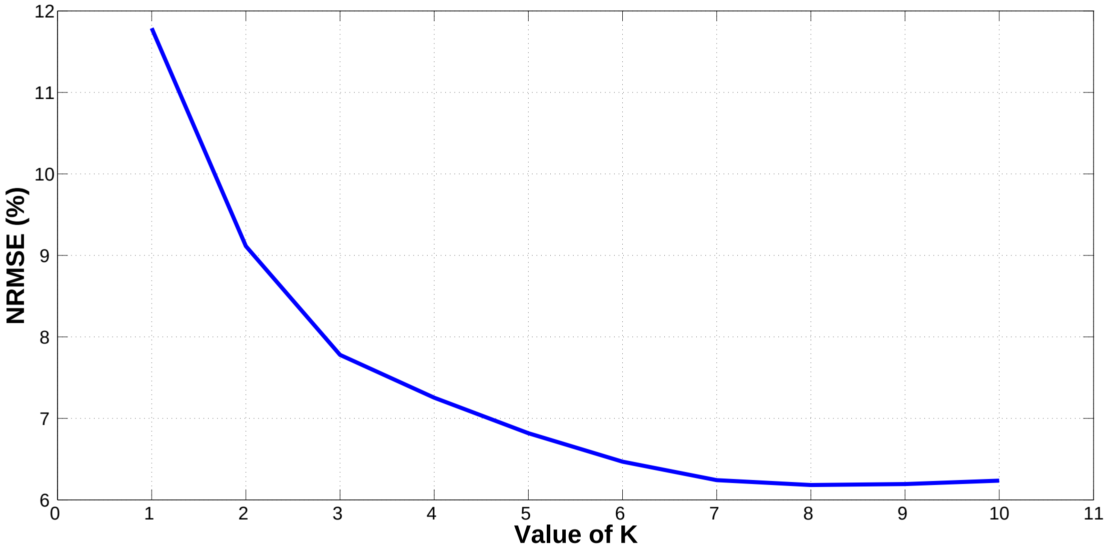


Figure 5.35: Average of Trial-Averaged-Val-NRMSE Per K for Best Subset (Asthenia)

5.10 Strain Prediction

5.10.1 Greedy forward selection for ‘Strain’

The Greedy Forward Selection was described in Sections 5.5.4.1 and 5.6.1. The same procedure is used applied for ‘Strain’ prediction. Figures 5.36, 5.37 depict the average of trial-testing-NRMSE obtained for MLR and KNNR.

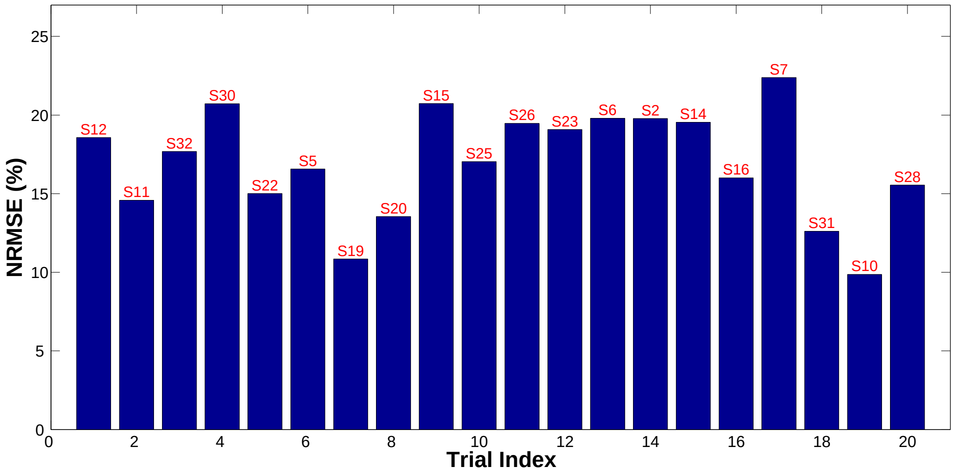


Figure 5.36: Trial-Testing-NRMSE for the best selected feature subset (MLR-Strain-Greedy). S labels are features subsets in Table E.9

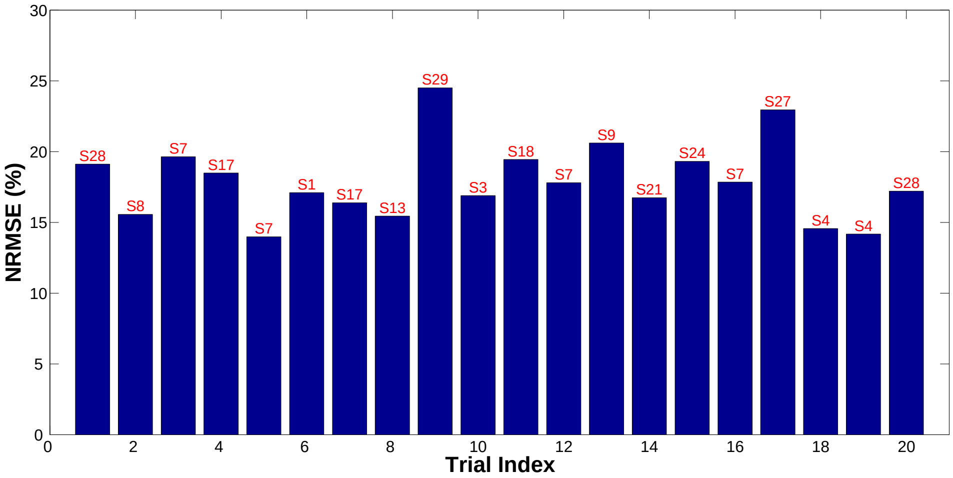


Figure 5.37: Trial-Testing-NRMSE for the best selected feature subset (KNNR-Strain-Greedy). S labels are features subsets in Table E.9

Table E.9 in the Appendix defines the feature subsets S1, S2, etc. referred to in these graphs. Each trial uses the best feature-subset found for that trial, and it may be different for each trial.

5.10.2 Exhaustive search for 'Strain'

The Exhaustive Search procedure was explained in Section 5.5.4.2 and 5.6.2. After discarding the 10 features with lowest Pearson correlation with Strain as described in Section 5.4, the rest of the features were pre-selected for an 'exhaustive search'. The ten selected features are the first ten as presented in Table (5.12), i.e. F4, F14, F20, F11, F6, F13, F18, F5 ,F19 and F1. Box-plots for the 20 best performing feature-subsets, as selected from the 1023 possible feature-subsets, are presented in Figure 5.38 (for MLR) and 5.39 (for KNNR).

The labels S1, S2, etc. refer to feature subsets as defined in Table E.10 in the appendix. The 'top twenty' best performing subsets are those with the lowest average of trial-averaged validation-NRMSE. The feature subset with the lowest average of trial-averaged validation-NRMSE is feature-subset 'S11' for MLR and 'S28' for KNNR. According to Table E.10 , these subsets contain features, i.e. F4, F14, F20, F5 and F19 and F1 (MLR) and F4, F14, F6, F5 and F1 (KNNR). Figure 5.40 and 5.41 show the value of average of trial-testing-NRMSE that was obtained for each of the 1023 different feature-subsets.

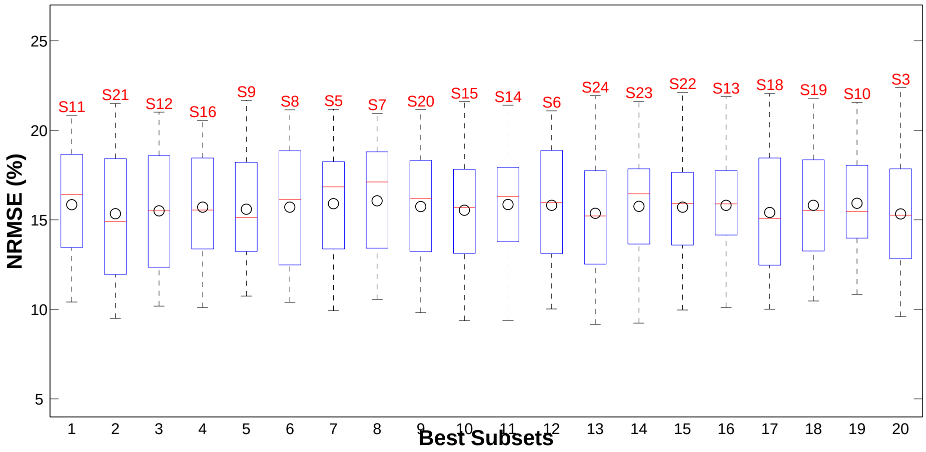


Figure 5.38: Average of Trial-Testing-NRMSE for the best 20 selected feature subset (MLR-Strain-EXHAUSTIVE). S labels are feature subsets in Table E.10. ('o':average, '-':median, '+':outliers)

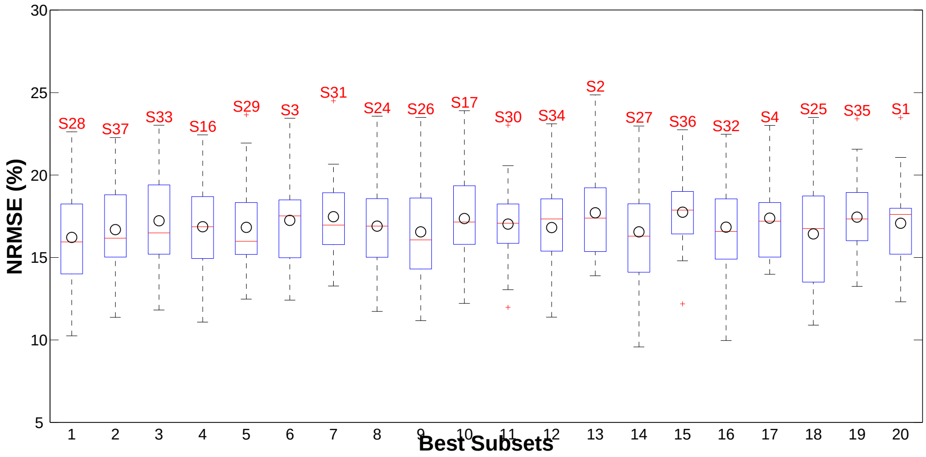


Figure 5.39: Average of Trial-Testing-NRMSE for the best 20 selected feature subset (KNNR-Strain-EXHAUSTIVE). S labels are feature subsets in Table E.10. ('o':average, '-':median, '+':outliers)

These curves demonstrate that improvement and deterioration in the prediction accuracy of either method, compared with what is achieved using all ten features, can be achieved by feature selection. The scope for improvement is rather small, i.e. of the order of about 2% like 'G', 'R', 'B' and 'A'.

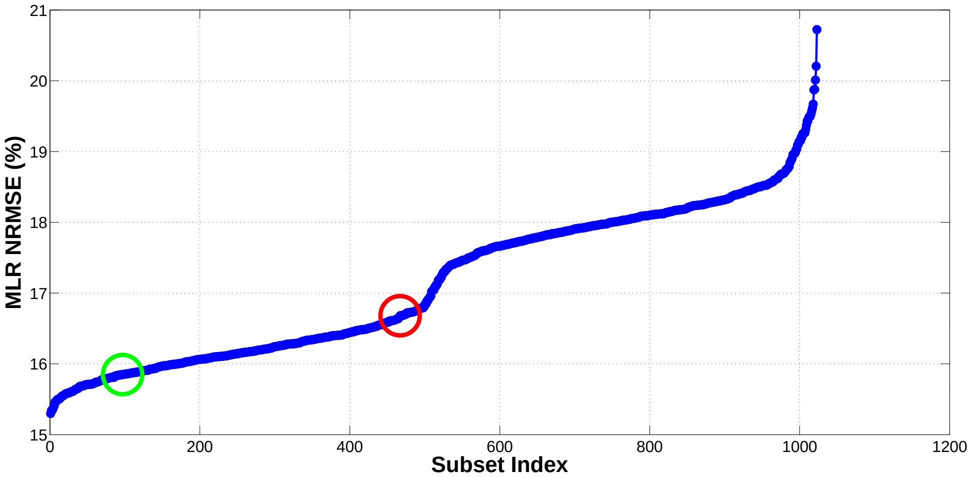


Figure 5.40: Average of Trial-Testing-NRMSE for all subset of features. green circle best subset (KNNR-Strain-EXHAUSTIVE). (green circle: best subset and red circle: 10 features)

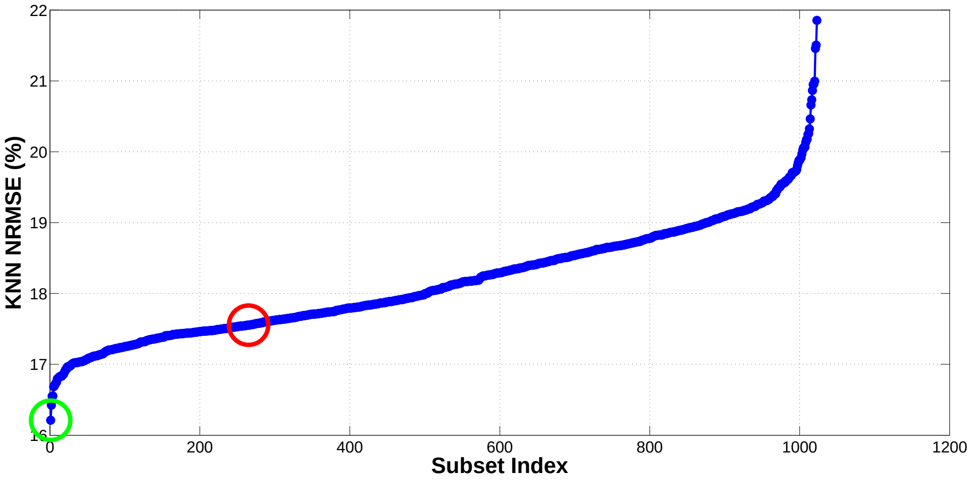


Figure 5.41: Average of Trial-Testing-NRMSE for all subset of features (KNNR-Strain-EXHAUSTIVE). (green circle: best subset and red circle: 10 features)

5.10.3 Optimum K for Strain prediction by KNNR

The K selection for procedure was described in Section 5.6.3. Figure 5.42 shows the trial-averaged validation-NRMSE in each trial for K, 1 to 10 when the best subset (S28) was found for 'Strain' prediction. The best subset with the lowest averaged-averaged validation-NRMSE has K=8.

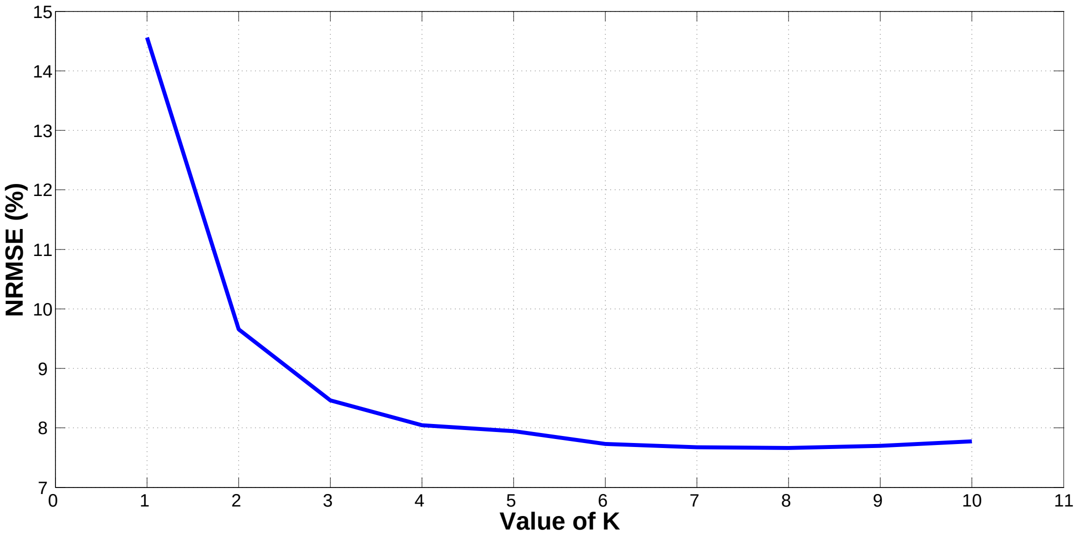


Figure 5.42: Average of Trial-Averaged-Val-NRMSE Per K for Best Subset (Strain)

5.11 Exhaustive search compared with greedy forward selection

'Exhaustive search' is computationally much more expensive than 'Greedy Forward Selection' but it explores the effect of all subsets on the prediction models and is therefore more likely to find the best subset. With greedy forward selection, not all possible subsets are considered, and the best subset may be different in each trial because of differences in the randomized choices of training and testing subjects. Greedy search can be an appropriate low-complexity method for finding out the best subset, but the best approach is likely to be exhaustive search if the computation can be afforded.

5.12 Comparison between MLR and KNNR

The performance of the MLR and KNN techniques, both with and without feature selection, are now compared for predicting GRBAS scores objectively. The standard deviation of the error may be investigated to estimate the stability of the MLR and KNNR models. In table 5.13 for 'all features' the average of trial-averaged-testing-NRNSE and the standard deviation of trial-testing-NRNSE and the confidence interval and the confidence level are presented. With confidence interval, we can say about the confidence of the results for the the true percentage of the population. For example, for MLR the average of trial-testing-NRNSE and the standard deviation of trial-testing-NRNSE is about 21.44% and 3.55 respectively with 95% confidence limits at 19.88% and 22.99%. For KNNR, the corresponding average of trial-testing-NRNSE and the standard deviation of trial-testing-NRNSE is about 23.93% and 4.10% respectively with 95% confidence limits at 22.0% and 25.59% . The same applied to 'R', 'B', 'A' and 'S'.

	Mean-Test-Error %		STD-Test-Error %		Lower-Bound %		Upper-Bound %	
GRBAS	MLR	KNNR	MLR	KNNR	MLR	KNNR	MLR	KNNR
G	21.44	23.93	3.55	4.10	19.88	22.02	22.99	25.59
R	19.76	19.92	3.94	4.74	18.03	18.03	21.49	22.18
B	21.27	21.72	3.19	2.87	19.87	20.70	22.67	23.23
A	15.93	15.83	2.32	2.44	14.90	14.76	16.94	16.90
S	16.97	17.52	3.11	3.37	15.60	16.10	18.33	19.06

Table 5.13: Comparison Between MLR And KNNR for All twenty Features

Table 5.14 depicts the comparison between MLR and KNNR when the best obtained subset by 'exhaustive search' was used for the GRBASE prediction.

	Mean-Test-Error %		STD-Test-Error %		Lower-Bound %		Upper-Bound %	
GRBAS	MLR	KNNR	MLR	KNNR	MLR	KNNR	MLR	KNNR
G	20.94	23.30	3.79	3.47	19.28	21.78	22.61	24.83
R	17.02	18.96	3.33	4.21	15.56	16.84	18.49	20.54
B	19.34	20.19	2.69	2.50	18.16	19.09	20.52	21.28
A	13.75	15.68	2.70	2.99	12.56	14.37	14.93	16.99
S	15.84	16.21	3.18	2.88	14.45	14.94	17.24	17.47

Table 5.14: Comparison Between MLR And KNNR for the Best Subset

For all GRBAS component with 'all-features' and the 'best-subset' the confidence limits overlap, is concluded that no statistically significant difference between the models is observed at the 95% confidence level.

5.13 Conclusions

This chapter has introduced two machine learning methods for objectively predicting GRBAS scores, together with feature selection methods that may be used to improve the resulting prediction models. The machine learning models were first implemented without feature selection. Then, after surveying feature selection in general, two feature selection methods were implemented and studied in detail to determine their effectiveness for reducing the number of dimensions in the feature measurements and for reducing the prediction error of regression models which use selected feature subsets.

The two main types of feature selection methods, i.e. ‘filter’ and ‘wrapper’ methods were investigated. A filter method which selects features based upon Pearson correlation was investigated. Wrapper methods which select features based upon the performance of a regression algorithm were also investigated. ‘Exhaustive search’ and ‘greedy forward selection’ wrapper methods were implemented and compared. Results show that feature selection can improve the prediction results, but not dramatically. There is not statistically significant difference between the two prediction models. No single feature can completely characterise any of the GRBAS parameters for all voice disorders. The best subset for predicting each GRBAS component using MLR and KNNR (‘exhaustive search’) was found to be as

1. Grade:
    - MLR: Shimmer, CPP Min, CSID, API, MEPF, Mean CPP F0 STD
    - KNNR: Shimmer, CPP Min, API, MEPF, HNR, RMMEPF
2. Roughness
    - MLR: Shimmer, CPP STD, MEPF, API, STD EPF
    - KNNR: Shimmer, STD EPF
3. Breathiness
    - MLR: CPP Max, MEPF, API
    - KNNR: CPP Max, MEPF, API, STD EPF
4. Asthenia
    - MLR: Shimmer, CPP, CPP Min, CPP Max, MEPF, Min L/H, API
    - KNNR: Shimmer, CPP Min, CPP Max, STD L/H, API
5. Strain
    - MLR: Shimmer, CPP Min, CSID, MEPF, Mean CPP F0 STD, API (MLR)
    - KNNR: Shimmer, CPP Min, RMMEPF, MEPF, API

The feature selection was made for the subset with the lowest the average of trial-averaged-validation-NRMSE. The next chapter evaluates the performance of objective methods (MLR and KNNR with and without feature selection) and perceptual analysis against the ‘reference GRBAS scores’.

Chapter 6

Results and Evaluation

6.1 Introduction

The focus of this chapter is to evaluate the objective GRBAS scoring system, developed using the techniques investigated in this thesis, by comparing the scores it produces with those produced by the trained SLTs. Ideally a comparison with the scores produced by different SLTs would be preferred, but this was not possible. The philosophy of this evaluation is to assume that some average of the available scores is the 'gold standard' reference score for each GRBAS component, for each subject. The performance of the objective system is judged against these reference scores. Further evaluation is clearly necessary and is discussed as a proposal for further work.

The observed GRBAS scores of individual SLT and the predicted GRBAS scores of objective systems (MLR and KNNR) have been compared against the 'reference' GRBAS scores for the same number of subjects in each trial. Various approaches were discussed in Chapter 3 for trying to obtain the most reliable 'reference' GRBAS scores. We used the term 'gold standard' colloquially for these reference scores, though perhaps the term is a little misleading as there could never be an absolute undisputed standard set of scores. A method of deriving 'gold standard' reference scores taking into account the reliability and consistency of scorers was proposed in Chapter 3, though Chapter 5 is based on unweighted averages from five scorers being considered as the 'reliable' GRBAS scores. Further work will be based on the reference scores produced by Chapter 3.

Evaluation of the objective systems (MLR and KNNR) has been performed in two cases. Firstly, we considered the case where measurements of all 20 features, as discussed in Chapter 4, are used to make predictions objectively. Secondly, we considered the case where feature selection is applied to identify the best subset of these twenty features, and this best subset is used for the prediction. The evaluations were based on the average of‘trial testing-NRMSE’ (normalized RMS) error between objectively predicted scores and ‘reference’ scores. This error is obtained from 20 trials, each with a different randomised selection of 80 training subjects and 22 ‘set-aside’ testing subjects. Based on the reference scores, SLTs can find out how their own scores compare with the average and the researcher may gain some idea of whether the objective scoring system is getting similar scores by objective means.

The objective scoring system is critically dependent on the ‘digital signal processing’ (DSP) algorithms used to measure the features considered indicative of voice quality. The study of such algorithms is a mature research topic, and there is much published and commercial software available. However, it became clear in Chapter 4 that strong discrepancies exist between the algorithms as provided by the best known published and commercial software packages. The ‘thesis software’ produced in this work attempts to improve this situation, but it cannot be concluded that the feature measurements used in Chapter 5 are totally reliable or fully indicative of voice quality as perceived by SLT scorers. Evaluation of some of the DSP discrepancies was given in Chapter 4, and these must be borne in mind when considering the evaluations in this chapter.

6.2 Objective scoring compared with SLT scores

6.2.1 MLR and KNNR for ‘Grade’ scoring

The observed GRBAS scores of individual SLT and the predicted GRBAS scores of objective systems (MLR and KNNR) have been compared against the ‘reference’ GRBAS scores (average of scores of five scorers) for the 22 ‘set-aside’ testing subjects in each trial. The evaluations were based on the average of‘trial testing-NRMSE’ (normalized RMS) error between objectively predicted scores and ‘reference’ scores. NRMSE is a percentage of the maximum GRBAS score of ‘3’. In Figure 6.1, zero represents the situation where the GRBAS score is equal to the reference score. For ‘Grade’ prediction using all 20 features, the average of trial-testing NRMSE was 21.44% for MLR and 23.93% for KNNR. Using the best obtained subset of features, the average of trial-testing NRMSE was around 20% for MLR and 23% for KNNR. Figure

6.1 shows the average of trial-testing NRMSE obtained for each of the five SLT scorers and the two objective (KNNR and MLR) voice quality assessment methods. The objective methods are plotted for the cases where all 20 features are used and where the best feature-subset (exhaustive search) is used. It may be seen that, for ‘Grade’, the objective methods deliver a higher error than is obtained for each of the SLT scorers. MLR seems to have performed better than KNNR, and MLR with the best subset performs marginally better than MLR with all 20 features. The discrepancy between the best performing objective method, i.e. ‘MLR’ with ‘best feature subset’ and the best scorer (SLT3) is about 8% which represents about one quarter of a GRBAS score in the range 0 to 3. The discrepancy between the best objective result (20.5%) and the worst SLT scorer result (19%) is about 1.5%.

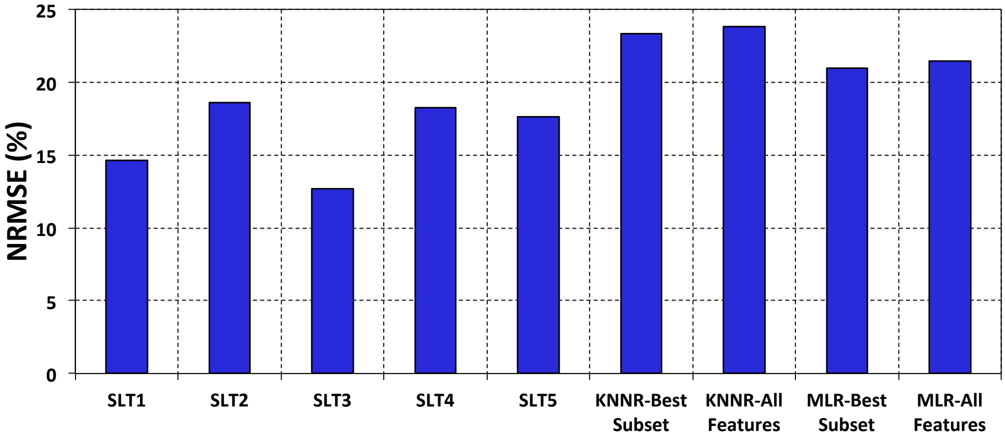


Figure 6.1: Average of Trial-Testing-NRMSE for subjective and objective ‘Grade’ scoring. The observed Grade scores of individual SLT and the predicted Grade scores of objective systems (MLR and KNNR) have been compared against the average of scores of five scorers for the 22 ‘set-aside’ testing subjects in each trial

6.2.2 MLR and KNNR for ‘Roughness’ scoring

Figure 6.2 plots the average of trial-testing NRMSE obtained for ‘Roughness’ for each of the five SLT scorers and the two objective (KNNR and MLR) methods without and with feature selection. The best objective result is again obtained using MLR with feature-selection. The error (17%) for this best objective result is roughly equal to the error for SLT5, significantly lower than the errors obtained for two of the scorers (SL2 and SLT4) , though it is significantly higher (by about 4%) than the errors for the remaining two SLT scores (SLT1 and SLT3).

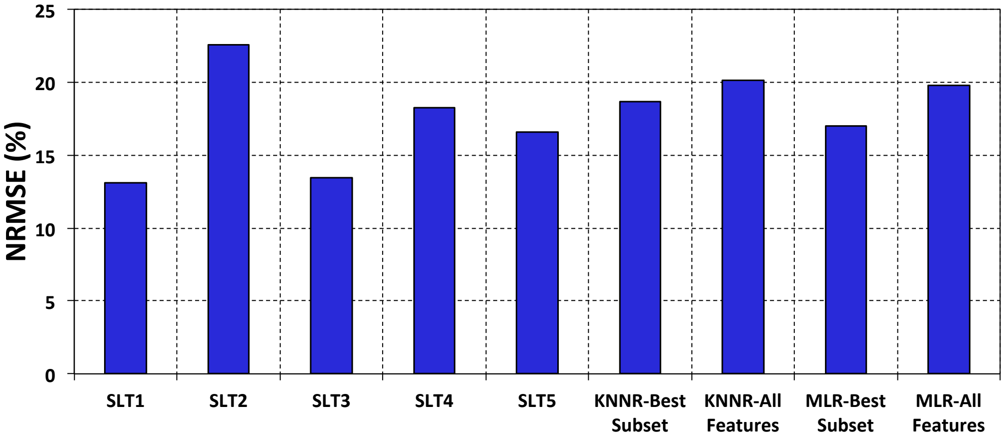


Figure 6.2: Average of Trial-Testing-NRMSE for subjective and objective 'Roughness' scoring. The observed 'Roughness' scores of individual SLT and the predicted Roughness scores of objective systems (MLR and KNNR) have been compared against the average of scores of five scorers for the 22 'set-aside' testing subjects in each trial

6.2.3 MLR and KNNR for 'Breathiness' scoring

Figure 6.3 plots the average of trial-testing NRMSE obtained for 'Breathiness' for each of the five SLT scorers and the two objective (KNNR and MLR) methods without and with feature selection. Once again, MLR with feature selection gives the best objective result though the error is higher than for all the five SLT scorers. This error is very close to that produced by one scorer (SLT4) but about 6% higher than the error for the best Breathiness SLT scorer.

6.2.4 MLR and KNNR for 'Asthenia' scoring

Figure 6.4 plots the average of trial-testing NRMSE obtained for 'Asthenia' for each of the five SLT scorers and the two objective (KNNR and MLR) methods without and with feature selection. The errors observed for both objective models (with and without feature selection) are lower than the errors for all five SLT scorers. MLR with feature selection works best and its error is about 2% lower than that obtained for the best performing SLT score (SLT3). It is about 8% better than the worst performing SLT(SLTs) for asthenia.

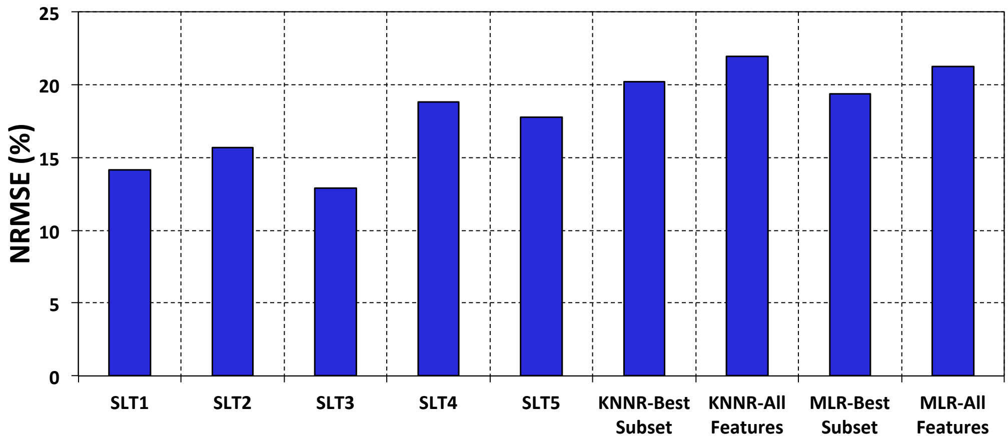


Figure 6.3: Average of Trial-Testing-NRMSE for subjective and objective 'Breathiness' scoring. The observed 'Breathiness' scores of individual SLT and the predicted Breathiness scores of objective systems (MLR and KNNR) have been compared against the average of scores of five scorers for the 22 'set-aside' testing subjects in each trial

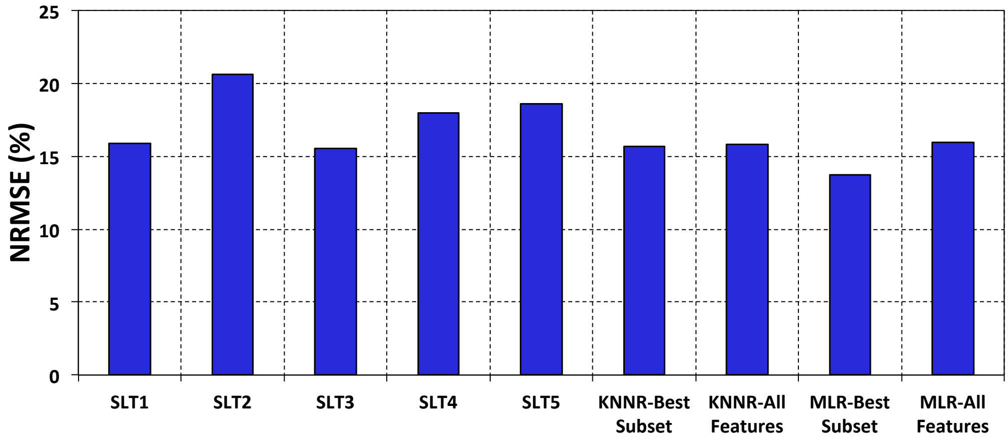


Figure 6.4: Average of Trial-Testing-NRMSE for subjective and objective 'Asthenia' scoring. The observed 'Asthenia' scores of individual SLT and the predicted Asthenia scores of objective systems (MLR and KNNR) have been compared against the average of scores of five scorers for the 22 'set-aside' testing subjects in each trial

6.2.5 MLR and KNN for 'Strain' scoring

Figure 6.5 plots the average of trial-testing NRMSE obtained for 'Strain' for each of the five SLT scorers and the two objective (KNNR and MLR) methods without and with feature selection.

MLR with feature selection is again the best objective method and, as with asthenia, it outperforms all five SLT scorers. The error is at least 1.5% lower than the lowest

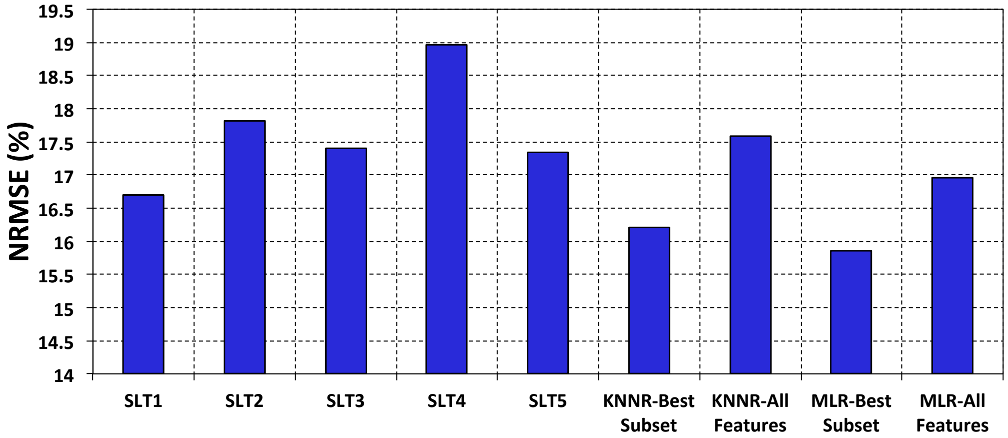


Figure 6.5: Average of Trial-Testing-NRMSE for subjective and objective 'Strain' scoring. The observed 'Strain' scores of individual SLT and the predicted Strain scores of objective systems (MLR and KNNR) have been compared against the average of scores of five scorers for the 22 'set-aside' testing subjects in each trial

SLT error (SLT1), and is 4% better than the worst performing SLT (SLT4). KNNR has slightly higher error than MLR. The errors for both objective models (MLR and KNNR), with feature selection, are lower than those for all five SLTs by MLR.

6.3 Conclusions

The performances of MLR and KNNR have been evaluated for predicting each GRBAS component with and without feature selection. The scores produced by these objective methods are compared with the individual scoring of the five SLTs whose averaged scores were used as the reference scores for training the prediction models. Although the argument may seem somewhat circular, we believe that comparing the objective predictions with the scores produced by each individual scorer gives some idea of how well the objective prediction is working.

It was found that MLR with feature selection was better than MLR without feature selection and KNNR with and without feature selection, for all five GRBAS components. It was also found that MLR with feature selection gives scores for 'Asthenia' and 'Strain' which are closer to the reference scores than the scores given by all five individual SLT scorers. The best objective score for 'Roughness' was closer than the scores given by two SLTs, roughly equal to the score of one SLT and worse than the other two SLT scores. The best objective scores for 'Breathiness' and 'Grade' were further from the reference scores than the scores produced by all five SLT scorers.

The worst 'MLR with feature selection' result has normalised RMS error which is only about 3% worse than the worst SLT scoring. We may conclude that results from objective scoring are encouraging and by no means completely at variance with the scoring that may be anticipated from traditional GRBAS scoring. It has been suggested [BPG04]that a combination of perceptual and objective scoring may have useful clinical applications and save some SLT time and effort.

It is relevant to point out that when this work started, there were only three scorers, and the additional scorings became available at quite a late stage when the approach to this work had already been formulated. Had five scorers been available from the beginning, we may have devised a different training and evaluation approach that, perhaps, reserved one set of scores for evaluation.

Chapter 7

Conclusions and Suggestions for Further Work

GRBAS scoring is universally used in Europe for voice quality assessment by subjective means in voice clinics. But there is not yet a universally accepted objective method for voice quality assessment where the outcome is expressed in terms of the well known and well understood GRBAS components. Research carried out to decide whether such a method is feasible has revealed many issues which led to several conclusions and some original insight.

Firstly, it was found that although much research has been carried out over many decades to find reliable digital signal processing (DSP) algorithms for measuring features considered indicative of voice quality, current algorithms cannot be considered reliable. There is much published and commercial software available for voice feature measurement. The most commonly referenced software packages are 'Praat' [Pa07], 'MDVP' [Kay96] and 'ADSV' [Kay96, AR05]. It became clear in Chapter 4, and it is widely reported in the literature [MCDB+09, HKŞ11, ARJ+10, AR09], that discrepancies exist between these software packages. Investigating these discrepancies is difficult even for the published algorithms in the 'Praat' package and is not possible for the commercial packages MDVP and ADSV. Also, these packages are designed as self-contained applications for users and currently do not make it possible for researchers to incorporate calls to their algorithms from 'voice analysis' software. The 'thesis software' produced in Chapter 4 attempts to improve this situation by implementing well known DSP techniques in an accessible form. Despite much effort, it cannot be concluded that the feature measurements developed in Chapter 4 are totally reliable or fully indicative of voice quality as perceived by SLT scorers. There is originality in the choice of techniques, for example choosing the ‘cross-correlation’ technique (see Section 4.3.5 over alternatives such as ‘autocorrelation’, frequency-domain and cepstral approaches for pitch-frequency detection and harmonic-to-noise measurements. All the ‘thesis software’ code is original. However the basic techniques used have been known for many years. The DSP development in this thesis was necessary to allow the ‘research hypothesis’ i.e. that computerized voice measurements can produce GRBAS scores comparable to subjectively assessed GRBAS scores, to be tested. We believe that the DSP development and the study of published and commercial packages have proved adequate for this purpose.

A second issue, considered in Chapter 3, emerged from an examination of the scoring produced by the five trained SLT scorers employed by Manchester Royal Infirmary. To facilitate the scoring process, a ‘GRBAS presentation and scoring package’ (GPSP) had to be developed. It was clear that some way of establishing a consensus between the five scorers and measuring the reliability and self-consistency of the scorers was needed before these scores were used for training machine learning techniques. A burning question was how to take measures of reliability and self-consistency into account when forming the consensus. It was concluded that this could be achieved, and the means of doing this are presented in Chapter 3. To allow self-consistency to be assessed the GPSP was adapted to request repeat scores for a number of randomly chosen subjects. The consistency of the scoring was analysed using correlation techniques and other more reliable techniques such as the ‘Cohen Kappa, ‘Fleiss Kappa’ and the ‘ICC’ measure. All these techniques are known, but a novel form of the Fleiss Kappa was presented in this chapter. The Cohen Kappa measures consistency between two scorers, whereas Fleiss Kappa measures consistency across many scorers, for example five scorers. Cohen Kappa is applicable to both categorical and ordinal data. However Fleiss Kappa, as published, is applicable only for categorical data. Chapter 3 proposes and evaluates a new form of Fleiss Kappa that is applicable to ordinal data such as GRBAS scores as well as categorical data. The insight gained in developing this new form of Fleiss Kappa also leads to further generalisations referred to as ‘Farideh Kappa’. Chapter 3 contains several original ideas including the design features of GPSP, the means of assessing self-consistency, the new form of Fleiss Kappa and the method of establishing a consensus across five scorers taking into account measures of self-consistency and inter-scorer consistency. It is concluded that these methods make better use of the scoring data than is achieved by straightforward averaging.

Further issues arose with the use, in Chapter 5, of machine learning to produce the objective predictions of GRBAS scores. The approach taken was regression rather than classification, and two simple approaches, i.e. ‘multiple linear regression’ (MLR) and ‘K nearest neighbours regression’ KNNR were investigated. More complex methods could have been chosen, but we believed that results obtained using the simpler approaches first would be indicative of what might be achieved by further research. As well as devising the means of training the chosen machine learning models, the means of testing the model had to be considered and implemented. The standard approach of setting aside randomly chosen subjects for testing, and conducting trials with different randomisations was adopted to try to make best use of the limited database with 102 scored subjects. Using this approach it was straightforward to train and test MLR and KNNR models using DSP measurements of twenty ‘voice quality’ features as listed in Chapter 4. A small complication was the need to find, using cross-validation, an appropriate value of K for KNNR as part of the training. Ten of the voice quality features were measured by the thesis software and the other ten were obtained from the commercial software package ‘ADSV’. The testing results obtained using all 20 features allowed us to conclude that GRBAS scores were being predicted reasonably, with a mean square error, over all trials, equivalent to about half of a GRBAS score on the scale 0 to 3. The error was computed with reference to a straightforward average of all five scorers, since the final results of Chapter 3 were not available when these tests were run. For most of this work, only 3 scorers were available, though we were able to update the averages when the extra two scorers became available.

To try to improve the results summarised above, feature selection was investigated. It was found that the number of features could be reduced from twenty to ten without significant loss of accuracy. This may be done either by ‘filtering’ according to Pearson correlation measures between individual features and the GRBAS scores, or using ‘greedy backward selection’ as a ‘wrapper’ method around the chosen MLR or KNNR prediction model. In fact, the simpler filtering approach was found to be satisfactory. Finally, both ‘greedy forward selection’ and an ‘exhaustive search’ procedure were implemented with the aim of improving the prediction accuracy by eliminating the chances of over-fitting and confusing the model with extraneous information. It was concluded that the improved performance of ‘exhaustive search’ over ‘greedy forward selection’ merited its selection for the final investigations. A ten-fold cross validation approach must be used to implement both ‘greedy forward selection’ and ‘exhaustive search’ before they are tested.

Results obtained using MLR and KNNR both with and without feature selection are presented in Chapter 6 together with a summary of the scores obtained by each of the five individual scorers. It was found that MLR with feature selection was better than MLR without feature selection and KNNR with and without feature selection, for all five GRBAS components. It was also found that MLR with feature selection gives scores for Asthenia and Strain which are closer to the reference scores than the scores given by all five individual SLT scorers. The best objective score for 'Roughness' was closer than the scores given by two SLTs, roughly equal to the score of one SLT and worse than the other two SLT scores. The best objective scores for breathiness and grade were further from the reference scores than the scores produced by all five SLT scorers.

The worst 'MLR with feature selection' result has normalized RMS error which is only about 3% worse than the worst SLT scoring. We may conclude that results from objective scoring are encouraging and by no means completely at variance with the scoring that may be anticipated from traditional GRBAS scoring. It has been suggested [HKŞ11] that a combination of perceptual and objective scoring may have useful clinical applications and save some SLT time and effort.

It was concluded that the proposed scheme based on measuring acoustic features and training prediction models can be helpful in assisting clinicians in their tasks of detecting voice abnormality. Two helpful aspects are:

1. Having the means of identifying the voice features that may be best for GRBAS prediction.

2. Having a method for obtaining a consensus of 'reliable GBBAS scores'.

It is useful for SLTs to be able to evaluate their scoring performance and compare it with the outputs of objective prediction models. Although, the objective system produced in this investigation may not be considered as accurate as it could be made given further research, it has the considerable advantage of being reproducible. Given the same data it will always produce the same result.

We believe that enhancements may be achieved by optimising the DSP algorithms, employing more SLTs for the GRBAS scoring and creating a larger database of scored subjects for training and evaluation. It may be observed that the objective results obtained are worse for the attributes 'G' and 'B' and better for the attributes 'R' ,'A' and 'S'. It would be useful to investigate why this is the case.

This thesis has considered whether and how voice quality can be objectively assessed according to the GRBAS scale. It does not aim to replace perceptual analysis but the methodolgy can be useful method in giving feedback to the clinicians and participants in clinics. In order to validate these results, further experiments with different databases must be carried out.

One highly original aspect of the work in this thesis is the use of the voice quality database set up at the Manchester Royal Infirmary by Chai G. Although we believe the GPSP scoring system was helpful in the scoring of this database, the credit for establishing it is due to Chai himself.

7.1 Suggestions for further Work

In addition to various suggestions for further research made in the previous chapters of this thesis, the following topics are highlighted as useful investigations:

1. MDVP, ADSV and 'thesis software' DSP algorithms have be used in this thesis for measuring features that are believed to be indicative of voiced quality. The Praat software package has been investigated and used to compare the measurements obtained. Clearly the success of an objective voice assessment package will be dependent on the choice of features, the definition of the parameters extracted from them, and the accuracy with which the features are measured. Despite the maturity of this research field, fundamental DSP research papers are still being published about feature measurement [FH09, MA07, FGHD+09] and other recent papers [AWA09, HKŞ11] have compared the accuracy and results obtained. We believe there is still much fundamental work to do in this area, and that the thesis software presented can be improved, extended and further evaluated.

2. We have taken the Euclidean distance between the k feature measurements of the new subject and the corresponding feature measurements of the data-base subject in KNNR. It is even possible to vary the definition of 'distance' dynamically according to the data and the confidence in the scores.

3. It would be useful to apply the proposed methodology to a different database such as that referred to in the publication by [VCOAAL+13]. Unfortunately, this database has not been scored by trained SLT therapists, only by one untrained voice researcher. It would be useful to employ our SLT scorers to score the voice recordings in this database and then use it to further train our system. This would allow a direct comparison with the results published in [VCOAAL+13].

4. The evaluations in Chapter 6 were carried out with reference to averaged scores rather than the reliability weighted average (Gold standard) scores developed in Chapter 3. The reasons for this were explained in the Page 215. The evaluations should now be repeated using the 'gold-standard' scores. This point is also mentioned in the conclusions of chapter 6 (last paragraph).

5. The measurements used for the system investigated were mostly extracted from sustained vowels, though the 'thesis software' allows connected speech, i.e. spoken sentences, to be used also. More work can now be carried out using feature measurements made from the voiced components of connected speech as identified by voiced/unvoiced detection.

6. Other features of speech, for example rate of change of jitter and shimmer, should be investigated enhancing GRBAS score prediction.

7. The effects of the age, sex, regional or international accent and state of mind of patients being assessed should be considered especially if connected speech is to be used for GRBAS prediction. The voice quality assessment of children presents formidable difficulty which has already been identified and needs to be addressed in further research. Questions arise, such

   (a) Which acoustic features are most indicative of voice abnormality in children?

   (b) How can voice recordings best be obtained from children?

8. One application of GRBAS measurement is the 'longitiudal study' of patients over periods of time. This raises interesting problems and opportunities such as how to maintain the consistency of repeated measurements, and whether mobile technology, perhaps mobile phones even, could be used for such monitoring.

9. Some very difficult questions remain, which have not been addressed in this thesis. For example, from the outset we decided to use objective measurements to emulate the assessment by SLT scorers. This will give a different emphasis to features that may not even be noticeable to human listeners. Therefore, it may be asked whether perceptual models should be built into the computer software. Also, SLT therapists have the advantage of seeing their patients, experiencing their discomfort and basing their diagnoses on subtle and perhaps indefinable cues that the software defined in this thesis would not even know how to look for.

This highlights again the need to make better use of connected speech for voice assessment. As well as the voiced segments of connected speech, the voiced-unvoiced transitions, and other characteristics are certain to contain diagnostic information. The use of connected speech for voice quality assessment is an issue that has been raised by many researchers, but, until now, remains relatively unexplored.

Bibliography

[Dr99] Dr.Speech. Dr. Speech software. `http://http://www.drspeech.com/`, 1999. [Online; accessed 19-September-2015].

[Pa07] Paul Boersma and David Weenink . Praat: doing phonetics by computer. `http://www.fon.hum.uva.nl/praat/`, 2007. [Online; accessed 19-June-2015].

[Re92] ReliaSoft Corporation. Chapter 4: Multiple Linear Regression Analysis . `http://reliawiki.org/index.php/Multiple_Linear_Regression_Analysis`, 1992. [Online; accessed 19-June-2015].

[ADS] ADSV Manual. A Division of PENTAX Medical Company 3 Paragon Drive Montvale, NJ 07645-1725 USA. [Accessed: 28-June-2015].

[AF94] Shaheen N Awan and Michael L Frenkel. Improvements in estimating the harmonics-to-noise ratio of the voice. *Journal of Voice*, 8(3):255–262, 1994.

[AR05] Shaheen N Awan and Nelson Roy. Acoustic prediction of voice type in women with functional dysphonia. *Journal of Voice*, 19(2):268–282, 2005.

[AR06] Shaheen N Awan and Nelson Roy. Toward the development of an objective index of dysphonia severity: a four-factor acoustic model. *Clinical linguistics & phonetics*, 20(1):35–49, 2006.

[AR09] Shaheen N Awan and Nelson Roy. Outcomes measurement in voice disorders: application of an acoustic index of dysphonia severity.

Journal of Speech, Language, and Hearing Research, 52(2):482–499, 2009.

[ARJ+10] Shaheen N Awan, Nelson Roy, Marie E Jetté, Geoffrey S Meltzner, and Robert E Hillman. Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: comparisons with auditory-perceptual judgements from the cape-v. *Clinical linguistics & phonetics*, 24(9):742–758, 2010.

[Aro90] Arnold E Aronson. Importance of the psychosocial interview in the diagnosis and treatment of functional voice disorders. *Journal of Voice*, 4(4):287–289, 1990.

[AWA09] Ofer Amir, Michael Wolf, and Noam Amir. A clinical comparison between two acoustic analysis softwares: Mdvp and praat. *Biomedical Signal Processing and Control*, 4(3):202–205, 2009.

[B+00] Martin Bland et al. *An introduction to medical statistics.* Number Ed. 3. Oxford University Press, 2000.

[BB12] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[BF85] William D Berry and Stanley Feldman. *Multiple regression in practice*. Number 50. Sage, 1985.

[BKG+96] Steven Bielamowicz, Jody Kreiman, Bruce R Gerratt, Marc S Dauer, and Gerald S Berke. Comparison of voice analysis systems for perturbation measurement. *Journal of Speech, Language, and Hearing Research*, 39(1):126–134, 1996.

[Bla] Bland JM. How do I analyse observer variation studies? `http://wwwusers.york.ac.uk/~mb55/meas/observer.pdf`. [Accessed: 28-April-2015].

[BO00] Ronald J Baken and Robert F Orlikoff. *Clinical measurement of speech and voice*. Cengage Learning, 2000.

[Boe93] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, pages 97–110. Amsterdam, 1993.

[Boe09] Paul Boersma. Should jitter be measured by peak picking or by waveform matching? *Folia Phoniatrica et Logopaedica*, 61(5):305–308, 2009.

[BPG04] Tarika Bhuta, Linda Patrick, and James D Garnett. Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice*, 18(3):299–304, 2004.

[BS92] Leo Breiman and Philip Spector. Submodel selection and evaluation in regression. the x-random case. *International statistical review/revue internationale de Statistique*, pages 291–319, 1992.

[BW04] P Boersma and D Weenink. Praat manual. version 4.2. 17. *University of Amsterdam, Phonetic Sciences Department, Amsterdam, The Netherlands*, 2004.

[C$^+$60] Jacob Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[CD14] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014.

[C.G13a] C.Gadepalli, F.Jalalinajafabadi, F.Ascott, B.Cheetham, JJ.Homer. Inter & intra rater consistency in GRBAS scoring. http://britishlaryngological.org/event/cutting-edge-laryngology, 2013. [Accessed: 28-June-2013].

[C.G13b] C.Gadepalli, F.Jalalinajafabadi, F.Ascott, B.Cheetham, JJ.Homer. Voice burden in teachers. http://britishlaryngological.org/event/cutting-edge-laryngology, 2013. [Accessed: 28-June-2013].

[CHMA86] DG Childers, DM Hicks, GP Moore, and YA Alsaka. A model for vocal fold vibratory motion, contact area, and the electroglottogram. *The Journal of the Acoustical Society of America*, 80(5):1309–1320, 1986.

[CKRT99] Daniel E Callan, Ray D Kent, Nelson Roy, and Stephen M Tasko. Self-organizing map for the classification of normal and disordered female voices. *Journal of Speech, Language, and Hearing Research*, 42(2):355–366, 1999.

[CL06] Janina K Casper and Rebecca Leonard. *Understanding voice problems: A physiological perspective for diagnosis and treatment*. Lippincott Williams & Wilkins, 2006.

[Coh68] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[CTPB$^+$00] Paolo Campisi, Ted L Tewfik, Elaine Pelland-Blais, Murad Husein, and Nader Sadeghi. Multidimensional voice program analysis in children with vocal cord nodules. *Journal of Otolaryngology-Head & Neck Surgery*, 29(5):302, 2000.

[Dav79] Steven B Davis. Acoustic characteristics of normal and pathological voices. *Speech and language: advances in basic research and practice*, 1:271–335, 1979.

[Deb] Deborah J. Rumsey. from Statistics For Dummies, 2nd Edition. `http://media.wiley.com/product_data/excerpt/85/04709110/0470911085-37.pdf`. [Accessed: 28-July-2015].

[DHZS02] Chris Ding, Xiaofeng He, Hongyuan Zha, and Horst D Simon. Adaptive dimension reduction for clustering high dimensional data. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 147–154. IEEE, 2002.

[DNB$^+$02] Alireza Dibazar, S Narayanan, Theodore W Berger, et al. Feature analysis for automatic detection of pathological speech. In *Engineering in Medicine and Biology, 2002. 24th Annual Conference* *and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, volume 1, pages 182–183. IEEE, 2002.

[DW03] Om Deshmukh and Carol Espy Wilson. A measure of aperiodicity and periodicity in speech. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 1, pages I–385. IEEE, 2003.

[EI94] Massachusetts Eye and Ear Infirmary. Voice disorders database, version. 1.03 (cd-rom). *Lincoln Park, NJ: Kay Elemetrics Corporation*, 1994.

[FGHD$^+$09] Carlos Ferrer, Eduardo González, María E Hernández-Díaz, Diana Torres, and Anesto del Toro. Removing the influence of shimmer in the calculation of harmonics-to-noise ratios using ensemble-averages in voice signals. *EURASIP Journal on Advances in Signal Processing*, 2009:4, 2009.

[FGS$^+$07] Everthon Silva Fonseca, Rodrigo Capobianco Guido, Paulo Rogério Scalassara, Carlos Dias Maciel, and José Carlos Pereira. Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders. *Computers in Biology and Medicine*, 37(4):571–578, 2007.

[FH09] M Farrus and J Hernando. Using jitter and shimmer in speaker verification. *IET Signal Processing*, 3(4):247–257, 2009.

[Fit97] W Tecumseh Fitch. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America*, 102(2):1213–1222, 1997.

[FJP76] B Frøkjær-Jensen and S Prytz. Registration of voice quality. *Bruel and Kjaer Technical Review*, 3:3–17, 1976.

[Fle71] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[FLP81] L Fleiss, Bruce Levin, and Myunghee Cho Paik. The measurement of interrater agreement. In *In Statistical methods for rates and proportions (2nd ed.* Citeseer, 1981.

[FLP13] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[FSLGL+09] R Fraile, N Saenz-Lechon, JI Godino-Llorente, V Osma-Ruiz, and C Fredouille. Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex. *Folia phoniatrica et logopaedica*, 61(3):146–152, 2009.

[GE03] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[GHSS05] Lingyun Gu, John G Harris, Rahul Shrivastav, and Christine Sapienza. Disordered speech assessment using automatic methods based on quantitative measures. *EURASIP Journal on Applied Signal Processing*, 2005:1400–1409, 2005.

[GLGVBV06] Juan Ignacio Godino-Llorente, Pedro Gomez-Vilda, and Manuel Blanco-Velasco. Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. *Biomedical Engineering, IEEE Transactions on*, 53(10):1943–1953, 2006.

[GLORSL+08] Juan Ignacio Godino-Llorente, Víctor Osma-Ruiz, Nicolás Sáenz-Lechón, Ignacio Cobeta-Marco, Ramón González-Herranz, and Carlos Ramírez-Calvo. Acoustic analysis of voice using wpcvox: a comparative study with multi dimensional voice program. *European Archives of Oto-Rhino-Laryngology*, 265(4):465–476, 2008.

[GVB08] Adas Gelzinis, Antanas Verikas, and Marija Bacauskiene. Automated speech analysis applied to laryngeal disease categorization. *Computer Methods and Programs in Biomedicine*, 91(1):36–47, 2008.

[HAMB$^+$03] Yolanda D Heman-Ackah, Deirdre D Michael, Margaret M Baroody, Rosemary Ostrowski, James Hillenbrand, Reinhardt J Heuer, Michelle Horman, and Robert T Sataloff. Cepstral peak prominence: a more reliable measure of dysphonia. *Annals of Otology, Rhinology & Laryngology*, 112(4):324–333, 2003.

[HCE94] James Hillenbrand, Ronald A Cleveland, and Robert L Erickson. Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, 37(4):769–778, 1994.

[HH96] James Hillenbrand and Robert A Houde. Acoustic correlates of breathy vocal qualitydysphonic voices and continuous speech. *Journal of Speech, Language, and Hearing Research*, 39(2):311–321, 1996.

[Hil88] James Hillenbrand. Perception of aperiodicities in synthetically generated voices. *The Journal of the Acoustical Society of America*, 83(6):2361–2371, 1988.

[Hir81] Minoru Hirano. *Clinical examination of voice*, volume 5. Springer New York, 1981.

[HKŞ11] OĞUZ Haldun, Mehmet Akif Kiliç, and Mustafa Asım Şafak. Comparison of results in two acoustic analysis programs: Praat and mdvp. *Turk J Med Sci*, 41(5):835–841, 2011.

[HMWM66] Harry Hollien, Paul Moore, Ronald W Wendahl, and John F Michel. On the nature of vocal fry. *Journal of Speech, Language, and Hearing Research*, 9(2):245–247, 1966.

[Hol87] Harry Hollien. old voices: What do we really know about them? *Journal of voice*, 1(1):2–17, 1987.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Unsupervised learning*. Springer, 2009.

[IVL70] S Iwata and H Von Leden. Pitch perturbations in normal and pathologic voices. *Folia Phoniatrica et Logopaedica*, 22(6):413–424, 1970.

[JGA$^+$13] Farideh Jalalinajafabadi, Chaitanya Gadepalli, Frances Ascott, Jarrod Homer, Mikel Luján, and Barry Cheetham. Perceptual evaluation of voice quality and its correlation with acoustic measurement. In *Modelling Symposium (EMS), 2013 European*, pages 283–286. IEEE, 2013.

[Jia02] Yu Jiangsheng. Method of k-nearest neighbors. *Institute of Computational Linguistics, Peking University, China*, 100871, 2002.

[JJG$^+$97] Barbara H Jacobson, Alex Johnson, Cynthia Grywalski, Alice Silbergleit, Gary Jacobson, Michael S Benninger, and Craig W Newman. The voice handicap index (vhi) development and validation. *American Journal of Speech-Language Pathology*, 6(3):66–70, 1997.

[K$^+$95] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145, 1995.

[Kay96] KayPENTAX. A Division of PENTAX medical Company. `http://www.kaypentax.comf`, 1996. [Online; accessed 19-July-2015].

[KG03] Jody Kreiman and Bruce R Gerratt. Jitter, shimmer, and noise in pathological voice quality perception. In *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.

[KG05] Jody Kreiman and Bruce R Gerratt. Perception of aperiodicity in pathological voice. *The Journal of the Acoustical Society of America*, 117(4):2201–2211, 2005.

[KGA$^+$09] Gail B Kempster, Bruce R Gerratt, Katherine Verdolini Abbott, Julie Barkmeier-Kraemer, and Robert E Hillman. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, 18(2):124–132, 2009.

[KJ97] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[KK90] Dennis H Klatt and Laura C Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

[KKK90] W Bastiaan Kleijn, Daniel J Krasinski, and Richard H Ketchum. Fast methods for the celp speech coding algorithm. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 38(8):1330–1342, 1990.

[Kli82] RJ Klich. Effects of speech level and vowel context on intraoral air pressure in vocal and whispered speech. *Folia Phoniatrica et Logopaedica*, 34(1):33–40, 1982.

[KZP07] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. Emerging Artificial Intelligence Applications in Computer Engineering, 2007.

[LC11] Wei Liu and Sanjay Chawla. Class confidence weighted knn algorithms for imbalanced data sets. In *Advances in Knowledge Discovery and Data Mining*, pages 345–356. Springer, 2011.

[LCB82] Christy L Ludlow, David C Coulter, and Celia J Bassich. Relationships between vocal jitter, age, sex, and smoking. *The Journal of the Acoustical Society of America*, 71(S1):S55–S56, 1982.

[LK77] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[LKB00] Jonh Laver, RD Kent, and MJ Ball. Phonetic evaluation of voice quality. *Voice quality measurement*, pages 37–48, 2000.

[MA07] Peter J Murphy and Olatunji O Akande. Noise estimation in voice signals using short-term cepstral analysis. *The Journal of the Acoustical Society of America*, 121(3):1679–1690, 2007.

[MAT12] MATLAB. MATLAB GUI. `http://uk.mathworks.com/discovery/matlab-gui.html`, 2012. [Accessed: 28-June-2012].

[MB94] Reinhold Müller and Petra Büttner. A critical discussion of intraclass correlation coefficients. *Statistics in medicine*, 13(23-24):2465–2476, 1994.

[MBE10] Lindasalwa Muda, Mumtaj Begam, and I Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010.

[MCDB$^+$09] Youri Maryn, Paul Corthals, Marc De Bodt, Paul Van Cauwenberge, and Dimitar Deliyski. Perturbation measures of voice: a comparative study between multi-dimensional voice program and praat. *Folia Phoniatrica et Logopaedica*, 61(4):217–226, 2009.

[MDV] MDVP Manual. MDVP manual. Version 2.7.0, Kay Elemetrics Corporation, Lincoln Park, New Jersey, USA. [Accessed: 28-June-2015].

[MFW95] David Martin, James Fitch, and Virginia Wolfe. Pathologic voice type and the acoustic prediction of severity. *Journal of Speech, Language, and Hearing Research*, 38(4):765–771, 1995.

[MGS$^+$11] Claudia Manfredi, Andrea Giordano, Jean Schoentgen, Samia Fraj, Leonardo Bocchi, and Philippe Dejonckere. Validity of jitter measures in non-quasi-periodic voices. part ii: The effect of noise. *Logopedics Phoniatrics Vocology*, 36(2):78–89, 2011.

[MGS$^+$12] Claudia Manfredi, Andrea Giordano, Jean Schoentgen, Samia Fraj, Leonardo Bocchi, and Philippe H Dejonckere. Perturbation measurements in highly irregular voice signals: Performances/validity of analysis software tools. *Biomedical signal processing and control*, 7(4):409–416, 2012.

[MW96] Kenneth O McGraw and Seok P Wong. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30, 1996.

[MY01] Estella PM Ma and Edwin ML Yiu. Voice activity and participation profileassessing the impact of voice disorders on daily activities. *Journal of Speech, Language, and Hearing Research*, 44(3):511–524, 2001.

[Nat14] National Center for Voice and Speech. Electroglottography. http://www.ncvs.org/ncvs/tutorials/youngexp/howsee.html, 2014. [Accessed: 28-July-2015].

[OB90] Robert F Orlikoff and RJ Baken. Consideration of the relationship between the fundamental frequency of phonation and vocal jitter. *Folia Phoniatrica et Logopaedica*, 42(1):31–40, 1990.

[OSB^{+}89] Alan V Oppenheim, Ronald W Schafer, John R Buck, et al. *Discrete-time signal processing*, volume 2. Prentice hall Englewood Cliffs, NJ, 1989.

[PIGP^{+}93] William F Punch III, Erik D Goodman, Min Pei, Lai Chia-Shun, Paul D Hovland, and Richard J Enbody. Further research on feature selection and classification using genetic algorithms. In *ICGA*, pages 557–564, 1993.

[Rab77] Lawrence R Rabiner. On the use of autocorrelation analysis for pitch detection. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 25(1):24–33, 1977.

[Ras88] Jeffrey Lee Rasmussen. Evaluating outlier identification tests: Mahalanobis d squared and comrey dk. *Multivariate Behavioral Research*, 23(2):189–202, 1988.

[RB74] WJ Ryan and KW Burk. Perceptual and acoustic correlates of aging in the speech of males. *Journal of communication disorders*, 7(2):181–192, 1974.

[RMM02] RT Ritchings, M McGillion, and CJ Moore. Pathological voice quality assessment using artificial neural networks. *Medical engineering & physics*, 24(7):561–564, 2002.

[RMR07] Calyampudi Radhakrishna Rao, J Philip Miller, and Dabeeru C Rao. *Handbook of statistics: epidemiology and medical statistics*, volume 27. Elsevier, 2007.

[RMT^{+}04] Nelson Roy, Ray M Merrill, Susan Thibeault, Rahul A Parsa, Steven D Gray, and Elaine M Smith. Prevalence of voice disorders in teachers and the general population. *Journal of Speech, Language, and Hearing Research*, 47(2):281–293, 2004.

[RS11] LR Rabiner and RW Schafer. Digital speech processing. *The Froehlich/Kent Encyclopedia of Telecommunications 6 (2011): 237*, 258, 2011.

[Sam] Samuel Bakouch. Producing vowels and LPC. `http://cours.etudes.ecp.fr/c/MA2500/work/assig_10/Lab_3_Assignment_-_Samuel_Bakouch_1.pdf`. [Accessed:4-March-2016].

[Sam06] Hanan Samet. *Foundations of multidimensional and metric data structures*. Morgan Kaufmann, 2006.

[SBD05] François Severin, Baris Bozkurt, and Thierry Dutoit. Hnr extraction in voiced speech, oriented towards voice quality analysis. In *Proc. eUSiPcO*, volume 5, 2005.

[SC$^+$07] Kumara Shama, Niranjan U Cholayya, et al. Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology. *EURASIP Journal on Applied Signal Processing*, 2007(1):50–50, 2007.

[SCDB05] Ilse Smits, Piet Ceuppens, and Marc S De Bodt. A comparative study of acoustic voice measurements by means of dr. speech and computerized speech lab. *Journal of Voice*, 19(2):187–196, 2005.

[Sch95] Ronald C Scherer. Laryngeal function during phonation. *Diagnosis and treatment of voice disorders*, pages 86–104, 1995.

[SF79] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[SGG00] Joseph C Stemple, Leslie E Glaze, and Bernice K Gerdeman. *Clinical voice pathology: Theory and management*. Cengage Learning, 2000.

[She03] David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2003.

[SLGLORGV06] Nicolas Saenz-Lechon, Juan I Godino-Llorente, Víctor Osma-Ruiz, and Pedro Gomez-Vilda. Methodological issues in the development of automatic systems for voice pathology detection. *Biomedical Signal Processing and Control*, 1(2):120–128, 2006.

[SLORGL$^+$08] Nicolás Sáenz-Lechón, Víctor Osma-Ruiz, Juan Godino-Llorente, Manuel Blanco-Velasco, Fernando Cruz-Roldán, Julián D Arias-Londono, et al. Effects of audio compression in automatic detection of voice pathologies. *Biomedical Engineering, IEEE Transactions on*, 55(12):2831–2835, 2008.

[SM82] Steven J Schwager and Barry H Margolin. Detection of multivariate normal outliers. *The annals of statistics*, pages 943–954, 1982.

[SOA09] Darcio G Silva, Luís C Oliveira, and Mario Andrea. Jitter estimation algorithms for detection of pathological voices. *EURASIP Journal on Advances in Signal Processing*, 2009:9, 2009.

[ULTC12] SK Ueng, CM Luo, TY Tsai, and H Chang. Voice quality assessment and visualization. In *Complex, Intelligent and Software Intensive Systems (CISIS), 2012 Sixth International Conference on*, pages 618–623. IEEE, 2012.

[VCOAAL$^+$13] T Villa-Canas, JR Orozco-Arroyave, JD Arias-Londono, JF Vargas-Bonilla, and JI Godino-Llorente. Automatic assessment of voice signals according to the grbas scale using modulation spectra, mel frequency cepstral coefficients and noise parameters. In *Image, Signal Processing, and Artificial Vision (STSIVA), 2013 XVIII Symposium of*, pages 1–5. IEEE, 2013.

[VdBZDJ57] JW Van den Berg, JT Zantema, and P Doornenbal Jr. On the air resistance and the bernoulli effect of the human larynx. *The journal of the acoustical society of America*, 29(5):626–631, 1957.

[VG$^+$05] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.

[vHP93] Vincent J van Heuven and Louis C Pols. *Analysis and synthesis of speech: strategic research towards high-quality text-to-speech generation*, volume 11. Walter de Gruyter, 1993.

[VMJ96] Maunlio N Vieira, Fergus R McInnes, and Mervyn A Jack. Robust f 0 and jitter estimation in pathological voices. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 2, pages 745–748. IEEE, 1996.

[VS09] Miltiadis Vasilakis and Yannis Stylianou. Voice pathology detection based eon short-term jitter estimations in running speech. *Folia Phoniatrica et Logopaedica*, 61(3):153–170, 2009.

[VV98] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[Wag13] Isolde Wagner. A new jitter-algorithm to quantify hoarseness: an exploratory study. *International Journal of Speech Language and the Law*, 2(1):18–27, 2013.

[WCD$^+$04] AL Webb, PN Carding, IJ Deary, Kenneth MacKenzie, Nick Steen, and JA Wilson. The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Oto-Rhino-Laryngology and Head & Neck*, 261(8):429–434, 2004.

[WF05] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[WFC95] Virginia Wolfe, James Fitch, and Richard Cornell. Acoustic prediction of severity in commonly occurring voice problems. *Journal of Speech, Language, and Hearing Research*, 38(2):273–279, 1995.

[WS87] Virginia I Wolfe and Thomas M Steinfatt. Prediction of vocal severity within and across voice types. *Journal of Speech, Language, and Hearing Research*, 30(2):230–240, 1987.

[YGB82] Eiji Yumoto, Wilbur J Gould, and Thomas Baer. Harmonics-to-noise ratio as an index of the degree of hoarseness. *The journal of the Acoustical Society of America*, 71(6):1544–1550, 1982.

[Yiu02] Edwin ML Yiu. Impact and prevention of voice problems in the teaching profession: embracing the consumers' view. *Journal of Voice*, 16(2):215–229, 2002.

[YTGF99] Huang Yuan, Shian-Shyong Tseng, Wu Gangshan, and Zhang Fuyan. A two-phase feature selection method using both filter and wrapper. In *Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on*, volume 2, pages 132–136. IEEE, 1999.

[Zim94] Donald W Zimmerman. A note on the influence of outliers on parametric and nonparametric tests. *The journal of general psychology*, 121(4):391–401, 1994.

[ZJ08] Yu Zhang and Jack J Jiang. Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *Journal of Voice*, 22(1):1–9, 2008.

Appendix A

Ethical approval

NHS

National Research Ethics Service

North West 5 Research Ethics Committee - Haydock Park

NHS North West
Room 155 - Gateway House
Piccadilly South
Manchester
M60 7LP

Telephone: (0161) 237 2394 / 2152
Facsimile: (0161) 237 2383

17 November 2009

Mr J Homer
Consultant Head and Neck Surgeon / Otolaryngologist
Department of Otolaryngology – Head and Neck Surgery
Central Manchester University Hospitals NHS Foundation Trust
Manchester Royal Infirmary
Oxford Road
MANCHESTER
M13 9WL

Dear Mr Homer

Title of the Research Database:	**Voice quality database**
REC reference:	**09/H1010/65**

Thank you for your letter of 06 November 2009, responding to the Committee's request for further information on the above research database and for submitting revised documentation.

The further information has been considered on behalf of the Committee by Dr T Sprosen – Vice-Chair and Medical Researcher / Epidemiologist.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion of the above research database on the basis described in the application form and supporting documentation (as revised).

Duration of ethical opinion

The favourable opinion is given for a period of five years from the date of this letter and provided that you comply with the conditions set out in the attached document. You are advised to study the conditions carefully. The opinion may be renewed for a further period of up to five years on receipt of a fresh application. It is suggested that the fresh application is made 3-6 months before the 5 years expires, to ensure continuous approval for the research database.

Approved documents

The documents reviewed and approved at the meeting were: -

Document	*Version*	*Date*
Covering Letter		21 September 2009

REC application		21 September 2009
Protocol for Management of the Database	1	21 September 2009
Participant Information Sheet: Case	2	06 November 2009
Participant Information Sheet: Control	2	07 November 2009
Participant Consent Form	1	21 September 2009
The VoiSS - Voice Symptom Scale		
Response to Request for Further Information - Covering letter from Mr Homer		06 November 2009

Research governance

A copy of this letter is being sent to the R&D office responsible for Central Manchester University Hospitals NHS Foundation Trust.

Under the Research Governance Framework (RGF), there is no requirement for NHS research permission for the establishment of research databases in the NHS. Applications to NHS R&D offices through IRAS are not required as all NHS organisations are expected to have included management review in the process of establishing the database.

Research permission is also not required by collaborators at data collection centres (DCCs) who provide data under the terms of a supply agreement between the organisation and the database. DCCs are not research sites for the purposes of the RGF.

Database managers are advised to provide R&D offices at all DCCs with a copy of the REC application for information, together with a copy of the favourable opinion letter when available. All DCCs should be listed in Part C of the REC application.

NHS researchers undertaking specific research projects using data supplied by a database must apply for permission to R&D offices at all organisations where the research is conducted, whether or not the database has ethical approval.

Site-specific assessment (SSA) is not a requirement for ethical review of research databases. There is no need to inform Local Research Ethics Committees.

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees (July 2001) and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

After ethical review

Now that you have completed the application process please visit the National Research Ethics Service website > After Review

Here you will find links to the following: -

a) Providing feedback. You are invited to give your view of the service that you have received from the National Research Ethics Service and the application procedure. If you wish to make your views known please use the feedback form available on the website.

b) Annual Reports. Please refer to the attached conditions of approval.
c) Amendments. Please refer to the attached conditions of approval.

We would also like to inform you that we consult regularly with stakeholders to improve our service. If you would like to join our Reference Group please email referencegroup@nres.npsa.nhs.uk

09/H1010/65 **Please quote this number on all correspondence**

Yours sincerely

CAStokes

pp. **Dr Donal Manning**
Chair

E-mail: - cathie.stokes@northwest.nhs.uk

Enclosure:- Conditions of Ethical Approval

Copy to: - Alison Robinson
Research Information Officer
Research & Development Office
1st Floor, Postgraduate Medical Centre
Manchester Royal Infirmary
Oxford Road
MANCHESTER
M13 9WL

Central Manchester University Hospitals NHS
NHS Foundation Trust

Research & Development
1st Floor Post Graduate Centre
Manchester Royal Infirmary
Oxford Road
Manchester M13 9WL
Tel: 0161-276-3340
Fax: 0161-276-5766
Lorraine.Broadfoot@cmft.nhs.uk
http://intranet.xcmmc.nhs.uk/directorates/deptr
and

Mr Jarrod Homer
Consultant Head and Neck Surgeon/Otolaryngologist
Central Manchester Foundation Trust
Manchester Royal Infirmary and Christie Hospital
Oxford Road
Manchester
M13 9WL

Ref: R03021—-Ltr 2-HOMER

Dear Mr Homer

PIN: R03021 (Please quote this number in all future correspondence)
REC Reference: 12/NE/0305
Research Study: Voice Burden in General Population

Thank you for submitting the above study for approval.

We acknowledge that Central Manchester Foundation Trust has accepted the role of Research Governance Sponsor for this study.

We understand that this study is not adopted by the NIHR Portfolio.

I am pleased to confirm that the Research Office has now received all necessary documentation, and the Trust Director of Research & Innovation has given approval for the project to be undertaken. This approval is in relation to the documentation supplied to us below.

Approval is given subject to the attached conditions – please ensure you and all members of the research team are familiar with these before commencing your research.

Please note: You must tell your Divisional Research Manager -

- the date that you intend to start recruiting to this study AND
- the date on which the first participant is recruited/consented

The Trust aims for its research projects to recruit their first participant within 30 days of the recruitment start date. If you do not tell us your actual recruitment start date, we will use this approval date. This information is important for monitoring Trust recruitment performance for internal and external assessment.

I would like to take this opportunity to wish you well with your research.

Yours sincerely

LBroadfoot

Lorraine Broadfoot
Research Operations Manager

Date:15/08/12

cc. Mr Gadepalli

Documents Acknowledged/Approved

Document	Version Number / Reference	Date
NRES Approval Letter		13/08/2012
GP/Consultation Information Sheets	Letter to Firsway Health Centre	08/08/2012
Investigator CV	CV Mr Homer	02/04/2012
Letter of Invitation to Participant	Invitation Letter 1	02/04/2012
CV Mr Gadepalli		
CV Mr Dunlop		
Participation Information Sheet	3	30/08/2012
Protocol	Protocol 1	20/01/2012
Questionnaire – Occupational Voice Disorder Survey	2	08/08/2012
REC application		30/07/2012
Summary Synopsis	Flow Chart 1	20/01/2012

NHS

Health Research Authority

NRES Committee North East - County Durham & Tees Valley

Room 002
TEDCO Business Centre
Viking Industrial Park
Rolling Mill Road
Jarrow
Tyne & Wear
NE32 3DT

Telephone: 0191 4283545
Facsimile: 0191 4283432

07 September 2012

Mr Jarrod Homer
Consultant Head and Neck Surgeon/Otolaryngologist
Central Manchester Foundation Trust
Manchester Royal Infirmary and Christie Hospital
Oxford Road
Manchester
M13 9WL

Dear Mr Homer

Full title of study: **Voice burden in general population**
REC reference number: **12/NE/0305**

Thank you for your letter of 5th September 2012. I can confirm the REC has received the documents listed below as evidence of compliance with the approval conditions detailed in our letter dated 13 August 2012. Please note these documents are for information only and have not been reviewed by the committee.

Documents received

The documents received were as follows:

Document	*Version*	*Date*
Covering Letter		05 September 2012
Participant Information Sheet	V3	30 August 2012

You should ensure that the sponsor has a copy of the final documentation for the study. It is the sponsor's responsibility to ensure that the documentation is made available to R&D offices at all participating sites.

12/NE/0305 **Please quote this number on all correspondence**

Yours sincerely

Miss Hayley Jeffries
Committee Co-ordinator

E-mail: hayley.jeffries@nhs.net

Copy to: Ms Lynne Webster, Central Manchester University Hospitals NHS Foundation Trust

Mr Gadepalli Chaitanya, Senior Research Fellow, ENT, Central Manchester Foundation Trust

GM PC ReGrouP Greater Manchester Primary Care Research Governance Partnership NHS

NHS SalfoR+D
Acute & Primary Care Research

NHS SalfoR+D Director: Professor Bill Ollier
R&D Associate Director: Rachel Georgiou

SalfoR+D web address: http://www.nhssalfordrd.org.uk/
ReGrouP web address: http://www.gmregroup.nhs.uk/index.html

18th September 2012

Mr Jarrod Homer
Consultant Head and Neck Surgeon/Otolaryngologist,
Central Manchester Foundation trust
Manchester Royal Infirmary,
Oxford Road,
Manchester
M13 9WL

Dear Mr Homer

Study Title: Voice burden in general population
REC Reference: 12/NE/0305
R&D Reference: 2012/053

Thank you for forwarding all the required documentation for your study as above. I am pleased to inform you that your study has been registered with NHS SalfoR+D and has gained NHS R&D approval from the following NHS Trusts:

- Trafford PCT

All clinical research must comply with the Health and Safety at Work Act, www.hse.gov.uk and the Data Protection Act. http://www.hmso.gov.uk/acts

It is a legal requirement for Principal Investigators involved in Clinical Trials to have completed accredited ICH GCP training within the last 2 years. Please ensure that you provide the R&D Department with evidence of this (certificate for completing the course). A list of GCP training courses can be obtained from the R&D Office.

All researchers who do not hold a substantive contract with the Trust must hold an honorary research contract before commencing any study activities related to this approval. The 'Research Passport Application Form'. This can be obtained from web addresses: http://www.gmregroup.nhs.uk/researchers/passports.html and http://www.hope-academic.org.uk/academic/salfordrd/Research%20Passports.html This form should be completed and returned, with a summary C.V and recent (within 6 months) CRB to the address shown above.

It is a condition of both NRES and NHS R&D approval that participant recruitment data should be forwarded on a regular basis. Therefore, progress reports must be submitted annually to the main REC and copied to the R&D office until the end of the study. http://www.nres.npsa.nhs.uk/applications/after-ethical-review/annual-progress-reports/

Where clinical trials of investigational medicinal products are sponsored by Salford Royal NHS Foundation Trust or Salford Primary Care Trust, it is a condition of Trust approval that Chief Investigators submit quarterly progress reports (to include Annual Safety Reports at the appropriate time) to R&D. For clinical trials of investigational medicinal products hosted within Salford Royal NHS Foundation Trust and Salford Primary Care Trust, the local PI will be expected to submit bi-annual progress reports to R&D. It is also a condition of approval that delegated duties (as agreed within clinical trial agreements and trial delegation logs) are fulfilled by only those delegated to undertake a specific duty. This will be monitored by the Sponsor's Representative during routine monitoring of the trial. Persistent non-compliance with these requirements may result in removal of Sponsorship or Trust R&D Approval.

Any amendments to the study should also be notified and approval sought by Ethics Committee and R&D Department. ***Where Salford Royal NHS Foundation Trust or Salford Primary Care Trust is acting as Sponsor then amendments or changes MUST be discussed with the Sponsor prior to REC submission.*** On completion of the study you are required to submit a 'Declaration of End of Study' form to the main REC, which should also be copied and forwarded to the R&D office at the address shown above.

Any serious adverse events or governance issues related to the research must be notified to the R&D office.

Yours sincerely,

SGowland

Sue Gowland
R&D Manager

Cc Chaitanya Gadepalli

Appendix B

Data-base Definition

There were originally 105 voice examples provided by MRI. Three of these following examples were eliminated because of recording problems in the sustained vowel /a/.

1. Example: 20120501-1 (recording problem)
2. Example: 20110930-1 (recording problem)
3. Example: 20110831-1 (high amplitude)

The number of recording examples that was played out to the each of the SLTs for GRBAS scoring was different. As mentioned in Chapter 3, ‘intra-scorer consistency’ was investigated for each scorer using GRBAS re-scoring for about randomly selected 20 participants. The following examples were eliminated for the SLT specified to make the number of repeated examples to each SLT equal.

1. SLT1: 20111004-2
2. SLT2: 20111004-3
3. SLT4: 20110912-1, 20110913-1
4. SLT5: 20110913-1, 20110919-1, 20111006-2

Appendix C

‘Gold-standard’ Reference Scores

This Appendix presents a table of the ‘gold-standard’ reference scores obtained in Chapter 3 by applying equations (3.28) to (3.31) for ‘Grade’, and corresponding equations for ‘R’, ‘B’, ‘A’ and ‘S’. The table also presents the unweighted averages of the scores for each subject (avG, avR, avB, avA, avS). The scores are for subjects n in the range 1 to 102.

n	G	R	B	A	S	avG	avR	avB	avA	avS
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.5	0.4	0.0	0.0	0.0	0.4	0.4	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	1.0	0.8	0.2	0.0	0.0	1.0	0.8	0.2	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.2	0.2	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	1.5	1.6	0.8	0.4	1.0	1.4	1.4	0.6	0.4	1.0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	0.6	0.4	0.6	0.5	0.4	0.6	0.4	0.6	0.4	0.4

Table C.1: ‘Gold-standard’ ref scores compared with unweighted averages (PART 1).

n	G	R	B	A	S	avG	avR	avB	avA	avS
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20	0.4	0.4	0.0	0.1	0.0	0.4	0.4	0.0	0.2	0.0
21	0.3	0.3	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0
22	0.3	0.3	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0
23	0.6	0.6	0.0	0.0	0.0	0.6	0.6	0.0	0.0	0.0
24	1.2	0.8	0.0	0.3	1.3	1.2	0.8	0.0	0.2	1.2
25	0.2	0.2	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0
26	1.9	1.7	1.7	1.5	0.0	1.8	1.6	1.6	1.4	0.0
27	2.8	0.8	2.7	1.9	1.7	2.8	0.6	2.8	2.0	1.8
28	0.4	0.2	0.2	0.1	0.0	0.4	0.2	0.2	0.2	0.0
29	2.5	1.8	0.9	1.3	0.7	2.4	1.8	1.0	1.2	1.0
30	1.0	1.0	0.0	0.3	0.4	1.0	1.0	0.0	0.4	0.4
31	1.9	1.8	1.4	0.9	1.0	1.8	1.8	1.4	0.8	1.0
32	1.2	1.2	0.0	0.0	0.1	1.2	1.2	0.0	0.0	0.2
33	0.3	0.0	0.0	0.3	0.0	0.2	0.0	0.0	0.2	0.0
34	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
35	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
36	0.8	0.8	0.2	0.3	0.3	0.8	0.8	0.2	0.2	0.2
37	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
38	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
39	0.2	0.2	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0
40	2.2	2.2	0.0	0.3	1.4	2.2	2.2	0.0	0.2	1.4
41	1.0	0.5	0.6	0.0	0.4	1.0	0.6	0.6	0.0	0.4
42	2.0	0.6	1.7	1.4	0.9	2.0	0.6	1.8	1.4	0.8
43	2.2	0.0	2.3	1.8	0.5	2.2	0.0	2.2	1.8	0.6
44	0.2	0.2	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0
45	0.5	0.4	0.2	0.0	0.3	0.4	0.4	0.2	0.0	0.2
46	2.9	2.2	0.5	1.3	2.9	2.8	2.2	0.4	1.2	2.8
47	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
48	1.2	1.1	0.4	0.7	1.0	1.2	1.0	0.4	0.6	1.0
49	2.0	2.0	0.6	0.9	0.7	2.0	2.0	0.6	0.8	0.6
50	2.9	2.9	0.6	1.3	1.7	2.8	2.8	0.6	1.2	1.8
51	2.4	1.3	2.0	1.7	1.3	2.4	1.2	2.0	1.8	1.2
52	1.6	1.1	0.8	1.7	0.7	1.6	1.0	0.8	1.6	0.6
53	1.2	0.4	0.7	0.5	1.0	1.2	0.6	0.6	0.4	1.0
54	1.6	0.6	0.6	0.7	1.6	1.6	0.6	0.6	0.6	1.6
55	1.4	1.0	0.9	0.7	1.0	1.4	1.0	1.0	0.6	1.0
56	1.0	0.1	0.4	0.1	0.3	1.0	0.2	0.4	0.2	0.2
57	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
58	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
59	1.4	1.4	0.6	0.6	0.6	1.4	1.4	0.6	0.6	0.6
60	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table C.2: 'Gold-standard' ref scores compared with unweighted averages (PART 2).

n	G	R	B	A	S	avG	avR	avB	avA	avS
61	0.5	0.2	0.2	0.3	0.0	0.4	0.2	0.2	0.2	0.0
62	1.8	1.4	0.7	1.8	0.7	1.6	1.2	0.6	1.6	0.6
63	1.8	1.1	1.7	0.0	0.9	1.6	1.0	1.6	0.0	0.8
64	0.6	0.0	0.3	0.6	0.0	0.6	0.0	0.4	0.6	0.0
65	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
66	0.2	0.0	0.2	0.3	0.0	0.2	0.0	0.2	0.2	0.0
67	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
68	1.5	1.4	0.4	0.6	0.3	1.4	1.4	0.4	0.6	0.2
69	2.9	2.1	2.2	1.7	2.6	2.8	2.0	2.2	1.6	2.6
70	1.6	1.4	0.0	0.7	0.9	1.6	1.4	0.0	0.8	1.0
71	1.0	1.0	0.2	0.3	0.6	1.0	1.0	0.2	0.2	0.6
72	2.6	1.1	2.2	2.0	1.4	2.6	1.0	2.2	2.0	1.4
73	2.0	1.2	1.9	1.1	1.1	2.0	1.2	1.8	1.2	1.2
74	1.0	0.1	0.7	1.0	0.3	1.0	0.2	0.6	1.0	0.2
75	3.0	1.5	2.7	1.8	2.4	3.0	1.2	2.8	2.0	2.4
76	2.1	0.8	1.8	1.9	0.3	2.0	0.8	1.8	1.8	0.2
77	3.0	3.0	2.2	1.3	1.5	3.0	3.0	2.0	1.2	1.4
78	2.8	2.7	0.9	0.8	1.5	2.6	2.6	0.8	0.8	1.4
79	2.3	2.4	0.5	0.8	1.1	2.4	2.4	0.4	1.0	1.0
80	1.2	1.0	0.9	0.7	0.8	1.2	1.0	0.8	0.8	0.6
81	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
82	2.2	0.8	2.0	1.2	1.0	2.2	0.8	2.0	1.2	1.0
83	1.6	0.5	1.7	1.4	0.3	1.6	0.4	1.6	1.4	0.2
84	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
85	2.2	2.1	1.8	1.3	1.2	2.2	2.0	1.6	1.2	1.4
86	0.9	0.4	0.7	0.0	0.0	0.8	0.4	0.6	0.0	0.0
87	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
88	2.0	0.6	2.0	1.7	0.3	2.0	0.4	2.0	1.6	0.2
89	2.5	0.8	0.2	0.8	2.5	2.4	0.6	0.2	0.8	2.4
90	0.2	0.0	0.0	0.2	0.0	0.2	0.0	0.0	0.2	0.0
91	2.5	2.2	0.5	0.3	1.9	2.4	2.0	0.4	0.2	1.8
92	0.9	0.9	0.0	0.1	0.4	0.8	0.8	0.0	0.2	0.4
93	2.6	2.6	0.8	0.5	0.8	2.6	2.6	0.8	0.4	0.8
94	1.4	1.4	0.4	0.3	0.2	1.4	1.4	0.4	0.4	0.4
95	0.2	0.2	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0
96	0.5	0.0	0.5	0.0	0.0	0.4	0.0	0.4	0.0	0.0
97	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
98	1.4	1.5	0.0	0.4	1.3	1.2	1.4	0.0	0.4	1.2
99	1.9	1.7	0.9	1.2	1.5	1.8	1.6	0.8	1.0	1.4
100	2.3	1.9	1.5	1.3	1.9	2.2	1.6	1.4	1.2	1.8
101	2.0	1.8	0.5	1.2	1.5	2.0	1.8	0.4	1.0	1.6
102	1.1	1.2	0.8	0.0	0.5	1.0	1.0	0.6	0.0	0.4

Table C.3: ‘Gold-standard’ ref scores compared with unweighted averages (PART 3).

Appendix D

Measuring Fundamental Frequency (FF) for artificial voiced speech

The purpose of this appendix is evaluating the ‘thesis software’ for fundamental frequency estimation using cross-correlation method for three sustained vowels (samples of artificial voiced speech) with known fundamental that were produced. These were produced by exciting an all-pole vocal tract model, with glottal pulse shaping and lip-radiation filtering, by a periodic series of discrete time impulses. Table D.1 in appendix D shows estimate of fundamental frequency for synthesised vowels, i.e. /a/, /o/ and /e/ over the range 120 Hz to 200 Hz as estimated by the thesis software.

Nominal-FF (Hz)	**th-FF-Vowel /a/**	**th-FF-Vowel /o/**	**th-FF-Vowel /e/**
120	119.84	119.83	119.84
130	130.08	130.08	130.09
140	140	139.99	140
150	149.98	149.99	150
160	159.77	159.77	159.78
170.	170.24	170.26	170.27
180	180	179.99	179.99
190	190.06	190.08	190.09
200	199.55	199.54	199.55

Table D.1: Comparison of Fundamental Frequency (FF) measurements for artificial voiced speech

The thesis measurements for three synthesised vowels is close to the nominal values of fundamental frequency. The maximum discrepancy is between the nominal fundamental frequency 200 Hz with thesis measurement for vowel /o/ with frequency 199.53 Hz that is about 0.23%. The small differences between the nominal and estimated values for three sustained vowels can arise is the mistaking of short term periodicity due to vocal tract resonances (formants) for the longer term pitch-cycle periodicity due to vocal cord vibration. The short term periodicity creates peaks in the cross-correlation function.

1. Poles were given radii of 0.987, 0.985, 0.919, and 0.929 with frequencies ±500, ±1200, ±2800 and ±3600 Hz respectively to emulate the phoneme /a/ [Sam, vHP93].

2. Poles were given radii of 0.985, 0.972, 0.919, and 0.909 with frequencies ±500, ±900, ±2250 and ±3200 Hz respectively to emulate the phoneme /o/ [Sam, vHP93].

3. Poles were given radii of 0.992, 0.99, 0.988, and 0.986 with frequencies ±500, ±1900, ±2100 and ±3400 Hz respectively to emulate the phoneme /e/ [Sam, vHP93].

Appendix E

Definitions of Feature Subsets used in Chapter 5

Subset	Feature Label
S1	F14-F2-F4-F5
S2	F14-F4-F5-F1-F20-F13
S3	F14-F4-F2-F20
S4	F20-F4-F1-F14-F19
S5	F4-F13
S6	F4-F13-F2-F19-F14
S7	F4-F13-F6-F2-F14
S8	F4-F14-F13-F1-F6-F2-F19
S9	F4-F14-F11-F13-F1-F19
S10	F4-F14-F1-F13-F20-F2
S11	F4-F14-F1-F2
S12	F4-F14-F6-F2-F5-F1-F20
S13	F4-F14-F6-F1-F5
S14	F4-F5-F6-F2-F1
S15	F4-F2-F13
S16	F4-F2-F14-F13-F20-F11-F5
S17	F4-F2-F14-F20
S18	F4-F2-F14-F19-F1-F5
S19	F4-F2-F14-F6
S20	F4-F2-F14-F6-F1
S21	F4-F2-F20-F14
S22	F4-F2-F20-F14-F19
S23	F4-F2-F5
S24	F4-F2-F5-F14-F19-F20-F11-F6-F13
S25	F4-F2-F5-F14-F1-F20
S26	F4-F2-F5-F11
S27	F4-F2-F1-F5-F14-F20
S28	F4-F2-F6
S29	F4-F2-F6-F13-F5-F14-F19-F1
S30	F4-F2-F6-F5-F1
S31	F4-F2-F6-F1
S32	F4-F2-F6-F1-F14-F11
S33	F4-F1-F14-F11-F2-F6-F19-F20
S34	F4-F1-F11
S35	F4-F1-F5-F19-F20-F14
S36	F4-F1-F6-F11
S37	F4-F1-F6-F11-F5-F14

Table E.1: Forward Selection subsets referred to in Figures 5.8 and 5.9 (Grade)

Subset	Feature Label
S1	F4-F14-F13-F1-F6
S2	F4-F14-F11-F1
S3	F4-F14-F11-F1-F5-F19
S4	F4-F14-F11-F1-F5-F2-F19
S5	F4-F14-F11-F1-F6
S6	F4-F14-F20-F13-F1-F5
S7	F4-F14-F20-F13-F1-F5-F19
S8	F4-F14-F20-F11-F1-F5
S9	F4-F14-F20-F11-F1-F5-F19
S10	F4-F14-F20-F5-F2
S11	F4-F14-F20-F5-F2-F19
S12	F4-F14-F20-F2
S13	F4-F14-F20-F1-F5
S14	F4-F14-F20-F1-F5-F19
S15	F4-F14-F20-F1-F5-F2
S16	F4-F14-F20-F1-F5-F6
S17	F4-F14-F20-F1-F2-F6
S18	F4-F14-F5-F2
S19	F4-F14-F5-F2-F19
S20	F4-F14-F5-F2-F6
S21	F4-F14-F2
S22	F4-F14-F2-F19
S23	F4-F14-F2-F6
S24	F4-F14-F1-F5
S25	F4-F14-F1-F5-F19
S26	F4-F14-F1-F5-F2
S27	F4-F14-F1-F5-F2-F6
S28	F4-F14-F1-F5-F6
S29	F4-F14-F1-F2-F6
S30	F4-F14-F1-F6
S31	F4-F11-F13-F1
S32	F4-F11-F2-F6
S33	F4-F11-F1
S34	F4-F11-F1-F6
S35	F4-F5-F2-F6
S36	F4-F2-F6
S37	F4-F1-F5-F2-F6
S38	F4-F1-F2-F6
S39	F4-F1-F6

Table E.2: Exhaustive search subsets referred to in Figures 5.10 and 5.11 (Grade)

Subset	Feature Label
S1	F14-F4-F12-F19-F7
S2	F20-F4
S3	F4
S4	F4-F14
S5	F4-F14-F12
S6	F4-F14-F20
S7	F4-F14-F1
S8	F4-F14-F1-F7-F12-F19
S9	F4-F12-F14
S10	F4-F12-F2-F7-F5
S11	F4-F12-F1-F5-F7
S12	F4-F12-F1-F2
S13	F4-F12-F7-F20-F1
S14	F4-F12-F7-F19-F1-F20-F14
S15	F4-F12-F7-F2
S16	F4-F20-F14-F12
S17	F4-F20-F5-F2-F14
S18	F4-F5
S19	F4-F5-F12
S20	F4-F5-F2-F14-F1
S21	F4-F5-F2-F12-F19
S22	F4-F5-F2-F12-F1
S23	F4-F5-F2-F7-F11-F14-F12
S24	F4-F5-F2-F7-F20-F12-F1-F14
S25	F4-F2
S26	F4-F2-F11
S27	F4-F2-F1-F11
S28	F4-F1-F11
S29	F4-F1-F7
S30	F4-F7-F14-F11-F1
S31	F4-F7-F14-F5
S32	F4-F7-F12-F19-F1
S33	F4-F7-F12-F2
S34	F4-F7-F12-F1
S35	F4-F7-F5-F1-F12
S36	F4-F7-F2
S37	F4-F7-F1
S38	F4-F7-F1-F12-F5
S39	F4-F7-F1-F5

Table E.3: Forward Selection subsets referred to in Figures 5.15 and 5.16 (Roughness)

Subset	Feature Label
S1	F4-F14
S2	F4-F14-F12-F5-F1-F7
S3	F4-F14-F11-F1-F7
S4	F4-F14-F20-F12-F5-F1-F7
S5	F4-F14-F20-F12-F5-F7
S6	F4-F14-F20-F12-F19-F1-F7
S7	F4-F14-F20-F12-F1-F7
S8	F4-F14-F20-F12-F7
S9	F4-F14-F20-F12-F7-F2
S10	F4-F12-F11-F19
S11	F4-F12-F5-F2
S12	F4-F12-F5-F1-F7
S13	F4-F12-F5-F1-F7-F2
S14	F4-F12-F5-F7
S15	F4-F12-F5-F7-F2
S16	F4-F12-F19
S17	F4-F12-F19-F5-F2
S18	F4-F12-F19-F1-F7
S19	F4-F12-F19-F1-F7-F2
S20	F4-F12-F19-F7
S21	F4-F12-F19-F7-F2
S22	F4-F12-F2
S23	F4-F12-F1-F7
S24	F4-F12-F7
S25	F4-F12-F7-F2
S26	F4-F11-F1
S27	F4-F11-F1-F2
S28	F4-F20
S29	F4-F20-F12-F5-F1-F7
S30	F4-F20-F12-F5-F1-F7-F2
S31	F4-F20-F12-F19
S32	F4-F20-F12-F1
S33	F4-F20-F12-F1-F7
S34	F4-F5
S35	F4-F5-F2
S36	F4-F5-F7
S37	F4-F2
S38	F4-F7

Table E.4: Exhaustive search subsets referred to in Figures 5.18 and 5.17 (Roughness)

Subset	Feature Label
S1	F13-F14-F2-F7-F5
S2	F13-F5
S3	F13-F5-F2
S4	F13-F5-F2-F10-F14-F4-F11-F7
S5	F13-F2-F5-F1
S6	F13-F2-F5-F1-F14
S7	F13-F1-F5-F7
S8	F13-F7-F4
S9	F14-F11-F2-F10-F5-F7-F13-F1
S10	F14-F4-F2-F11
S11	F11
S12	F11-F5-F13-F1-F7
S13	F11-F2
S14	F11-F2-F5-F13-F7
S15	F11-F2-F5-F1-F13-F10-F7-F14
S16	F11-F2-F1-F14-F5-F13-F4-F7
S17	F11-F2-F1-F5-F7-F13-F10
S18	F11-F2-F7-F14-F13
S19	F11-F2-F4-F13
S20	F11-F2-F4-F5-F14-F13
S21	F11-F1-F5-F13
S22	F11-F1-F7-F13
S23	F11-F1-F7-F13-F5
S24	F11-F1-F7-F13-F5-F2
S25	F11-F7-F1-F13-F5
S26	F11-F4-F13
S27	F5-F2-F14
S28	F5-F2-F11
S29	F1-F13-F7-F14-F11-F5-F2
S30	F1-F13-F7-F5-F2
S31	F1-F11-F7-F13-F5
S32	F4-F13-F11
S33	F4-F13-F5
S34	F4-F13-F5-F7
S35	F4-F13-F7-F5-F2-F20-F1-F14
S36	F4-F2-F13
S37	F4-F2-F5-F1-F13-F14-F20-F10

Table E.5: Forward Selection subsets referred to in Figures 5.22 and 5.23 (Breathiness)

Subset	Feature Label
S1	F13-F14-F5-F1
S2	F13-F14-F1-F7
S3	F13-F11-F5-F2-F1-F7
S4	F13-F11-F5-F2-F7
S5	F13-F11-F5-F1
S6	F13-F11-F5-F1-F7
S7	F13-F11-F2-F1-F7
S8	F13-F11-F2-F7
S9	F13-F11-F7
S10	F13-F11-F4-F5
S11	F13-F11-F4-F5-F2
S12	F13-F11-F4-F5-F1
S13	F13-F11-F4-F1-F7
S14	F13-F11-F4-F7
S15	F13-F20-F5-F1
S16	F13-F20-F5-F1-F10
S17	F13-F20-F5-F1-F7
S18	F13-F5-F2
S19	F13-F5-F2-F1
S20	F13-F5-F2-F1-F10
S21	F13-F5-F2-F1-F7
S22	F13-F5-F2-F10
S23	F13-F5-F2-F7
S24	F13-F5-F1
S25	F13-F5-F1-F7
S26	F13-F2-F7
S27	F13-F1-F7
S28	F13-F4
S29	F13-F4-F5
S30	F13-F4-F5-F2
S31	F13-F4-F5-F2-F1
S32	F13-F4-F5-F2-F1-F7
S33	F13-F4-F5-F1
S34	F13-F4-F5-F1-F7
S35	F13-F4-F1-F7

Table E.6: Exhaustive search subsets referred to in Figures 5.24,5.25 (Breathiness)

Subset	Feature Label
S1	F14-F13-F1-F5-F4-F16-F18-F11
S2	F14-F16-F1-F5-F4
S3	F14-F5-F13-F1-F4-F18-F16-F11-F20
S4	F14-F1-F5-F13
S5	F14-F1-F4-F11-F18-F5-F13-F20
S6	F14-F4-F11-F16
S7	F14-F4-F16-F11-F1-F13
S8	F14-F4-F16-F5-F1-F18
S9	F14-F4-F16-F1-F5
S10	F11-F13-F5
S11	F11-F14-F1-F4-F5-F13-F16-F18-F19-F20
S12	F11-F1-F5-F13-F4-F14-F18
S13	F11-F1-F4-F5-F20
S14	F11-F4-F1-F5-F14-F13-F19
S15	F20-F1-F5-F4-F11-F16-F13-F18-F14
S16	F4-F13-F11-F18-F5-F16-F1-F20
S17	F4-F13-F16-F1-F11
S18	F4-F13-F5
S19	F4-F13-F5-F20-F1-F14-F18
S20	F4-F13-F5-F18-F1-F16-F11-F20
S21	F4-F13-F19-F5-F14-F1-F18
S22	F4-F13-F19-F1-F16-F11
S23	F4-F13-F1-F14-F16
S24	F4-F14
S25	F4-F14-F5
S26	F4-F14-F5-F1-F19-F16-F13-F20
S27	F4-F14-F1
S28	F4-F14-F1-F13-F16
S29	F4-F11-F14-F1-F5-F16-F20-F13
S30	F4-F11-F5-F1-F19
S31	F4-F16-F20-F14-F1-F19-F13
S32	F4-F16-F5
S33	F4-F5-F1-F13-F16
S34	F4-F19-F13-F1-F14-F16
S35	F4-F1-F14-F13-F20-F11-F16-F5
S36	F4-F1-F14-F5-F20-F11
S37	F4-F1-F5-F13
S38	F4-F1-F5-F13-F18
S39	F4-F1-F5-F14-F13-F11-F18
S40	F4-F1-F5-F20

Table E.7: Forward Selection subsets referred to in Forward Selection subsets referred to in Figures 5.29 and 5.30 (Asthenia)

Subset	Feature Label
S1	F14-F5-F16-F1
S2	F4-F13-F5-F18-F1
S3	F4-F13-F5-F16-F1
S4	F4-F13-F5-F1
S5	F4-F13-F19-F16-F1
S6	F4-F13-F19-F5-F16-F1
S7	F4-F13-F19-F1
S8	F4-F14-F13-F16-F1
S9	F4-F14-F13-F5-F18-F1
S10	F4-F14-F13-F5-F16-F1
S11	F4-F14-F13-F5-F1
S12	F4-F14-F13-F19-F16-F1
S13	F4-F14-F13-F19-F1
S14	F4-F14-F13-F1
S15	F4-F14-F5-F18-F1
S16	F4-F14-F5-F16-F1
S17	F4-F14-F5-F1
S18	F4-F11-F13-F5-F16-F1
S19	F4-F11-F13-F5-F1
S20	F4-F11-F14-F13-F16-F1
S21	F4-F11-F14-F13-F5-F18-F1
S22	F4-F11-F14-F13-F5-F16-F18-F1
S23	F4-F11-F14-F13-F5-F16-F1
S24	F4-F11-F14-F13-F5-F1
S25	F4-F11-F14-F13-F19-F5-F1
S26	F4-F11-F14-F16-F1
S27	F4-F11-F20-F13-F5-F16-F18-F1
S28	F4-F11-F20-F14-F13-F16-F1
S29	F4-F11-F20-F14-F13-F5-F18-F1
S30	F4-F11-F20-F14-F13-F5-F1
S31	F4-F11-F5-F1
S32	F4-F20-F13-F16-F1
S33	F4-F20-F13-F5-F1
S34	F4-F20-F14-F13-F16-F1
S35	F4-F20-F14-F13-F5-F16-F1
S36	F4-F20-F14-F13-F5-F1
S37	F4-F19-F5-F16-F1

Table E.8: Exhaustive search subsets referred to in Figures 5.31 and 5.32 (Asthenia)

Subset	Feature Label
S1	F14-F4-F19
S2	F14-F4-F1-F18-F20-F6
S3	F4
S4	F4-F13
S5	F4-F13-F14-F19-F6-F5-F20
S6	F4-F13-F1-F5
S7	F4-F14
S8	F4-F14-F13
S9	F4-F14-F13-F11-F19
S10	F4-F14-F20
S11	F4-F14-F20-F19
S12	F4-F14-F20-F19-F11
S13	F4-F14-F20-F1-F11
S14	F4-F14-F5
S15	F4-F14-F19-F1
S16	F4-F14-F1-F13-F5
S17	F4-F14-F1-F11
S18	F4-F14-F1-F18-F11
S19	F4-F14-F1-F5
S20	F4-F11-F18-F14-F19-F5
S21	F4-F20
S22	F4-F20-F11-F14-F19
S23	F4-F20-F5-F6
S24	F4-F20-F1-F5-F14-F6
S25	F4-F18-F14-F19
S26	F4-F5-F14-F11
S27	F4-F5-F14-F1
S28	F4-F1
S29	F4-F1-F20
S30	F4-F1-F5-F14-F19-F20-F6
S31	F4-F1-F5-F19-F11-F14
S32	F4-F1-F19-F14-F20-F5

Table E.9: Forward Selection subsets referred to in Figures 5.36 and 5.37 (Strain)

Subset	Feature Label
S1	F4
S2	F4-F13
S3	F4-F14
S4	F4-F14-F11
S5	F4-F14-F11-F5-F19-F1
S6	F4-F14-F20-F11-F13-F5-F1
S7	F4-F14-F20-F11-F5-F19-F1
S8	F4-F14-F20-F11-F5-F1
S9	F4-F14-F20-F5
S10	F4-F14-F20-F5-F19
S11	F4-F14-F20-F5-F19-F1
S12	F4-F14-F20-F5-F1
S13	F4-F14-F20-F19
S14	F4-F14-F20-F19-F1
S15	F4-F14-F20-F1
S16	F4-F14-F20-F6-F5-F1
S17	F4-F14-F20-F6-F1
S18	F4-F14-F5
S19	F4-F14-F5-F19
S20	F4-F14-F5-F19-F1
S21	F4-F14-F5-F1
S22	F4-F14-F19
S23	F4-F14-F19-F1
S24	F4-F14-F1
S25	F4-F14-F6-F13-F5-F1
S26	F4-F14-F6-F18-F5-F1
S27	F4-F14-F6-F5
S28	F4-F14-F6-F5-F1
S29	F4-F14-F6-F1
S30	F4-F11-F1
S31	F4-F20-F1
S32	F4-F20-F6-F5
S33	F4-F20-F6-F5-F1
S34	F4-F1
S35	F4-F6-F18-F1
S36	F4-F6-F5
S37	F4-F6-F5-F1

Table E.10: Exhaustive search subsets referred to in Figures 5.38 and 5.39 (Strain)