Stress and Emotion Recognition in Natural Speech in the Work and Family Environments

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

Ling He

M.Eng. Signal Processing and Informatics
B.Eng. Telecommunication

School of Electrical and Computer Engineering
Science, Engineering and Technology Portfolio

RMIT University

November 2010
Contents

CONTENTS...........................................................................................................................I

LIST OF TABLES ................................................................................................................... V

LIST OF FIGURES ............................................................................................................... VIII

LIST OF ACRONYMS .......................................................................................................... XII

DECLARATION ..................................................................................................................... XIV

ACKNOWLEDGEMENTS ...................................................................................................... XV

ABSTRACT ............................................................................................................................ XVI

PATENTS AND PUBLICATIONS ......................................................................................... XVIII

CHAPTER 1. INTRODUCTION ............................................................................................ 1

1.1 Problem statements ........................................................................................................... 1

1.2 Thesis aims ....................................................................................................................... 2

1.3 Thesis scope .................................................................................................................... 3

1.4 Thesis original contributions .......................................................................................... 3

1.5 Major findings and conclusions ...................................................................................... 4

1.6 Thesis outline .................................................................................................................. 5

CHAPTER 2. REVIEW OF EXISTING METHODS OF STRESS AND EMOTION CLASSIFICATION IN SPEECH ......................................................................................... 8

2.1 What are stress and emotions? .......................................................................................... 8

2.1.1 Definitions of stress and emotion ............................................................................. 8

2.1.2 Types of emotions ...................................................................................................... 9

2.1.3 Simulated (acted), elicited and natural emotions ..................................................... 12

2.1.4 Data labeling (annotation) methods ......................................................................... 13

2.1.5 Existing emotional speech data collections .................................................................. 15

2.1.6 Language, culture and social environment based dependency of emotional expressions .................................................. 17

2.1.7 Speaker based dependency of emotional expressions .................................................. 18

2.1.8 Gender based dependency of emotional expressions .................................................. 18

2.2 How stress and emotion are expressed in speech? ....................................................... 19
2.2.1 Research rationale ......................................................................................................................19
2.2.2 Physiological basis .....................................................................................................................19
2.2.3 Acoustic properties of speech as indicators of stress or emotion ...............................................20

2.3 Why is it important to study emotional speech?....................................................................20
2.3.1 Applications of emotional speech technology .................................................................21
2.3.2 Benefits of the emotional speech research to the industry and society .......................................24

2.4 What are the existing methods of an automatic stress and emotion recognition in speech? ........................................................26
2.4.1 Levels of emotional speech analysis ..........................................................................................26
2.4.2 A general framework of the automatic emotion and/or stress recognition in speech ..........27
2.4.3 Existing methods of feature extraction in emotional speech analysis .........................................29
2.4.4 Existing feature selection methods for emotional speech ...........................................................34
2.4.5 Existing classification techniques for emotional speech ............................................................36

CHAPTER 3. GENERAL FRAMEWORK FOR STRESS AND EMOTION DETECTION IN SPEECH...................................................................................42
3.1 Speech data used in this study.................................................................................................42
3.1.1 The Speech Under Simulated and Actual Stress (SUSAS) data base .......................................42
3.1.2 The Oregon Research Institute (ORI) database ..........................................................................46

3.2 Stress and emotion classification system used in this study..................................................49
3.2.1 Pre-processing ............................................................................................................................51
3.2.2 Classification techniques ......................................................................................................52
3.2.3 Average Percentage of Identification Accuracy (APIA%) .........................................................58

CHAPTER 4. STRESS AND EMOTION CLASSIFICATION USING CLASSICAL FEATURES ........................................................................................................59
4.1 Classical source-filter model of speech production ...............................................................59
4.2 Calculating fundamental frequency (F0) and formants ..........................................................62
4.3 Calculating the Mel-Frequency Cepstral Coefficients (MFCCs) .........................................63
4.4 Estimation of the glottal waveform ..........................................................................................66
4.5 Glottal time and frequency domain features ...........................................................................68

4.6 Experiments and results........................................................................................................ ...70
4.6.1 Results of stress classification ...............................................................................................71
4.6.2 Results of emotion classification ..........................................................................................73

4.7 Conclusions and discussions ....................................................................................................74
4.7.1 Stress classification using classical features ...............................................................................74
4.7.2 Emotion classification using classical features ...........................................................................75

CHAPTER 5. NEW FEATURE EXTRACTION METHODS BASED ON SPEECH SPECTROGRAMS, WAVELET PACKETS AND EMPIRICAL MODE DECOMPOSITION ......................................................................................................77
5.1 New feature extraction methods based on speech spectrograms.................................77
5.1.1 Advantages of using speech spectrograms .................................................................77
5.1.2 Calculation of speech spectrograms ........................................................................78
5.1.3 Inspection of spectrograms for speech under stress and emotion ...........................78
5.1.4 New features based on speech spectrograms ............................................................80
5.1.5 Features extracted from speech sub-bands of spectrograms using different auditory scales (SS-CB-AE, SS-BARK-AE, SS-ERB-AE) ........................................................................80
5.1.6 Features extracted from speech spectrograms with anisotropic filtering and different auditory scales (SS-AF-CB-AE, SS-AF-BARK-AE, SS-AF-ERB-AE) ........................................................................84
5.1.7 Experiments and results for features SS-CB-AE, SS-BARK-AE, SS-ERB-AE, SS-AF-CB-AE, SS-AF-BARK-AE and SS-AF-ERB-AE ................................................................................85
5.1.8 Discussion of the results for SS-AF-ERB-AE, SS-AF-CB-AE, SS-AF-BARK-AE, SS-CB-AE, SS-BARK-AE and SS-ERB-AE features ........................................................................89
5.1.9 Stress and emotion recognition using log-Gabor filter analysis of spectrograms and optimal feature selection (SS-LGF-OFS, SS-ALGF-OFS and SS-SP-ALGF-OFS) ..............................................92
5.1.10 Experiments and results for SS-LGF-OFS, SS-ALGF-OFS and SS-SP-ALGF-OFS 97
5.1.11 Features extracted from speech spectrograms using sigma-pi units (SS-sigma-pi) ....103
5.1.12 Experiments and results for sigma-pi units (SS-sigma-pi) ........................................106
5.2 Features extracted from the wavelet packet arrays combined with a log-Gabor filter bank, averaging and an optimal feature selection (WP-ALGF-OFS) ................................................109
5.2.1 WP-ALGF-OFS feature extraction process ................................................................109
5.2.2 Calculation of the Wavelet Packet coefficients arrays .................................................110
5.2.3 Experiments and results for the WP-ALGF-OFS features ..........................................114
5.3 Features derived from the Empirical Mode Decomposition (EMD) combined with calculation of an averaged Renyi’s entropy (EMD-AER) ..........................................................117
5.3.1 Empirical Mode Decomposition (EMD) ..................................................................117
5.3.2 Experiments and results for the EMD-AER features ..................................................119

CHAPTER 6. NEW FEATURES BASED ON NONLINEAR MODELS OF SPEECH PRODUCTION ....................................................................................................................123

6.1 Introduction: nonlinear model of the glottal flow formation ........................................123
6.2 TEO based feature extraction .....................................................................................125
6.2.1 Teager Energy Operator ..............................................................................................126
6.2.2 Normalized TEO autocorrelation function ................................................................127
6.2.3 TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-P features .................128
6.2.4 Experiments and results for the TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-P features ..................................................................................................................131
6.2.5 Experiments and results for Stress and emotion classification using multilayer perceptron and probabilistic neural networks combined with TEO-PWP-S features ..........................................................135
6.3 Features based on the TEO contour ..........................................................................138
6.3.1 TEO-Contour .............................................................................................................138
6.3.2 Definitions of features based on the TEO-Contour .....................................................139
6.3.3 Experiments and results for the TEO-Contour-Mean, TEO-Contour-Median, TEO-Contour-Var, TEO-Contour-Min/Max, TEO-Contour-Diff and TEO-Contour-Diff2 ........................................140
6.4 Features derived from the spectral energy of glottal waveform (AUSEEG) and the speech (AUSEES) ............................................................................................................143
6.4.1 Definition of the AUSEEG and AUSEES features ......................................................143
6.4.2 Experiments and results using the AUSEEG and AUSEES features .......................146
List of Tables

Table 2-1 Emotion Labels in alphabetical ordering (based on Whissell [14], Plutchik [15] and Cowie [16]). ....................................................................................................... 11
Table 2-2 Description of emotional databases.............................................................................. 16
Table 3-1 A vocabulary of 35 words used in the SUSAS database............................................. 43
Table 3-2 Description of the Speech under Actual Stress datasets; H-high level stress, L-low level stress, N-neutral .......................................................................................... 45
Table 4-1 APIA% for stress classification using F₀, formants, MFCC, GP-T and GP-F features...................................................................................................................... 71
Table 4-2 TF statistics using ANOVA for stress (SUSAS data sets) using the GMM classifier; +denotes statistically significant difference (p<0.05), - denotes statistically insignificant difference (p>=0.05). ................................................................................... 72
Table 4-3 Confusion matrixes (%) for stress classification using MFCC features combined with the GMM classifier; H-high-level stress, L-low-level stress, N-neutral ...................................................................................................................... 72
Table 4-4 APIA% for emotion classification using F₀, formants, MFCC, GP-T and GP-F features...................................................................................................................... 73
Table 4-5 TF statistics using ANOVA for emotions (ORI data set) using the GMM classifier; +denotes statistically significant difference (p<0.05), - denotes statistically insignificant difference (p>=0.05). ................................................................................... 74
Table 4-6 Confusion matrixes (%) for emotion classification using MFCC features combined with the GMM classifier. ...................................................................................................................... 74
Table 5-1 Analysis frequency bands (critical bands, Bark scale and ERB scale). .................. 82
Table 5-2 APIA% for stress classification using spectrogram features without anisotropic filtering (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) and with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE, SS-AF-ERB-AE) ...................................................................................................................... 86
Table 5-3 Confusion table for stress classification using SS-ERB-AE feature combined with the GMM classifier; H-high level stress, L-low level stress, N-neutral ........... 88
Table 5-4 Confusion table for stress classification using SS-AF-ERB-AE feature combined with the GMM classifier; H-high level stress, L-low level stress, N-neutral ...................................................................................................................... 88
Table 5-5 APIA% for emotion classification using spectrogram features without anisotropic filtering (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) and with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE, SS-AF-ERB-AE). ....... 88
Table 5-6 Confusion table for the emotion classification using SS-ERB-AE and SS-AF-ERB-AE features combined with the GMM classifier; AE-Actual emotions, Ag-anger, Ax anxious, Dy-dysphoric, Ne-neutral, Ha-happy. ....................................... 89
Table 5-7 APIA% for stress classification for each of the 12 log-Gabor filters (SS-LGF-OFS) using the GMM and KNN classifiers.............................................................. 97
Table 5-8 APIA% for stress classification using SS-ALGF-OFS and SS-SP-ALGF-OFS features ...................................................................................................................... 99
Table 5-9 APIA% for emotion classification for each of the 12 log-Gabor filters (SS-LGF-OFS) using the GMM and KNN classifiers.............................................................. 100
Table 6-13 An example of a confusion table for stress classification using AUSEEG feature combined with the GMM classifier; H-high level stress, L-low level stress, N-neutral. ................................................................................................................ 146
Table 6-14 APIA% for emotion classification using AUSEEG and AUSEES features. 147
Table 6-15 An example of a confusion table for emotion classification using AUSEEG and AUSEES features combined with the GMM classifier ......................... 147
Table 7-1 Numbers of speech recordings from the ORI data base. The average length of a recording was 1.5 seconds; PD-parents of depressed adolescents, PND-parents of non-depressed adolescents. .................................................................................................................. 152
Table 7-2 The frequency bands and the corresponding bandwidth (Hz) for the TEO-PWP-S, TEO-PWP-G, AUSEEG and AUSEES methods using 11 kHz speech bandwidth...................................................................................................................................................... 154
Table 7-3 The APIA% for 7 emotions classification using TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG features and two classifiers: GMM and KNN; PD-parents of depressed adolescents, PND-parents of non-depressed adolescents. ................................................................................................................... 155
Table 7-4 Correct recognition rates for individual emotions using TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG features combined with the GMM classifier; Co-contempt, Ag-angry, Ax-anxious, Dy-dysphoric, Pl-pleasant, Ne-neutral, Ha-happy. ................................................................................................................................................... 156
Table 8-1 APIA% for stress and emotion classification using classical features: F0, formants, MFCC, GT and GF. .......................................................................................... 165
Table 8-2 APIA% for stress and emotion classification using TEO based features...... 165
Table 8-3 APIA% for stress and emotion classification using the AUSEEG and AUSEES features .................................................................................................................. 165
Table 8-4 TF-statistics from ANOVA for stress; + (p<0.05), - (p>=0.05); 1-F0, 2-Formants, 3-MFCC, 4-GT, 5-GF, 6-TEO-CB, 7-TEO-DWT, 8-TEO-WP, 9-TEO-PWP-S, 10-TWO-PWP-G, 11-AUSEEG-LB, 12-AUSEEG-CB, 13-AUSEEG-PWP, 14-AUSEES-LB, 15-AUSEES-CB, 16-AUSEES-PWP ......................................... 166
Table 8-5 TF-statistics from ANOVA for emotions; + (p<0.05), - (p>=0.05); 1-F0, 2-Formants, 3-MFCC, 4-GT, 5-GF, 6-TEO-CB, 7-TEO-DWT, 8-TEO-WP, 9-TEO-PWP-S, 10-TWO-PWP-G, 11-AUSEEG-LB, 12-AUSEEG-CB, 13-AUSEEG-PWP, 14-AUSEES-LB, 15-AUSEES-CB, 16-AUSEES-PWP ......................................... 167
Table 8-6 Emotion classification (APIA %) using AUSEEG-LB/GMM within selected band.......................................................................................................................... 172
Table 8-7 Confusion table for different stress levels based on MFCC and TEO-PWP-S features and GMM classifier .................................................................................. 173
Table 8-8 Confusion table for different emotions based on MFCC and AUSEEG-LB features and GMM classifier .................................................................................. 173
Table 8-9 Emotion classification (APIA %) using AUSEEG/GMM within selected frequency bands. ........................................................................................................... 175
Table 9-1 A ranking list of all feature/classifier combinations tested in this thesis using the same sets of data (SUAS for stress and ORI for emotions). Darker-grey - new nonlinear methods introduced in the thesis; lighter-gray - new linear methods introduced in the thesis. Feature that are not highlighted were previously used in literature. ........................................................................................................... 178
List of Figures

Figure 2-1 Areas and bounders of the 3-D emotion space (adapted from Dietz and Lang [13]) ................................................................. 10
Figure 2-2 The general flowchart of the speech stress and emotion recognition and classification system including two stages: training and testing (classification) .... 27
Figure 2-3 Feature selection using the filter approach .................................................. 34
Figure 2-4 Feature selection using the wrapper approach .............................................. 34
Figure 3-1 A general flowchart of the automatic stress or emotion recognition system. 50
Figure 3-2 An example of a multilayer perceptron neural network (MLPNN) structure used in the training and testing stages for five emotions classification (happy, anxious, neutral, angry and dysphoric) in speech ............................................. 55
Figure 3-3 An iterative training of the multilayer perceptron neural network (MLPNN) ................................................................. 56
Figure 4-1 Classical model of speech production assuming laminar flow of air through the vocal folds................................................................. 60
Figure 4-2 Summary of the source-filter model of speech production ............................ 62
Figure 4-3 A flowchart of the fundamental frequency estimation method .................... 63
Figure 4-4 Time and frequency domain representation of the speech production process. Speech samples s[n] are represented in the time domain as a convolution of the glottal, vocal filter and lip radiation impulse responses. In the frequency domain the speech z-transform S(z) is given as a product of the glottal, vocal filter and lip radiation transfer functions. ........................................................................ 66
Figure 4-5 Schematic of the glottal waveform indicating the time intervals corresponding to the “open” and “closed” phases; T denotes the glottal period ................. 68
Figure 4-6 APIA% for stress classification using F0, formants, MFCC, GP-T and GP-F features combined with the GMM classifier ................................................ 71
Figure 4-7 APIA% for stress classification using F0, formants, MFCC, GP-T and GP-F features combined with the KNN classifier .................................................. 72
Figure 4-8 APIA% for emotion classification using F0, formants, MFCC, GP-T and GP-F features ....................................................................................... 73
Figure 5-1 Examples of spectrograms for vowel a in the word “break” pronounced by the same parson; soft speech (a), neutral speech (b), speech under low level stress (c) and speech under high level stress (d) ........................................... 78
Figure 5-2 Examples of spectrograms for voiced parts of sentences pronounced with different emotions. Note that each spectrogram corresponds to a different sentence ....................................................................... 79
Figure 5-3 Features generation from the auditory frequency bands of spectrograms using different auditory scales (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) ............... 83
Figure 5-4 Feature extraction method using auditory frequency bands of anisotropic filtering spectrograms (SS-AF-CB-AE, SS-AF-BARK-AE and SS-AF-ERB-AE). 85
Figure 5-5 APIA% for stress classification using spectrogram features without anisotropic filtering (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) and with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE and SS-AF-ERB-AE) combined with the GMM classifier ........................................ 87
Figure 5-6 APIA% for stress classification using spectrogram features without anisotropic filtering (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) and with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE and SS-AF-ERB-AE) combined with the KNN classifier. 87

Figure 5-7 APIA% for emotion classification using spectrogram features without anisotropic filtering (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) and with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE, SS-AF-ERB-AE). 89

Figure 5-8 Features generation using single log-Gabor filter (SS-LGF-OFS). 95

Figure 5-9 Features generation using averaged outputs from log-Gabor filters (SS-ALGF-OFS). 96

Figure 5-10 Features generation using spectrogram patches and log-Gabor filters (SS-SP-ALGF-OFS). 96

Figure 5-11 APIA% for stress classification for each of the 12 log-Gabor filters (SS-LGF-OFS) and the GMM classifier; S-scale, O-Orientation. 98

Figure 5-12 APIA% for stress classification for each of the 12 log-Gabor filters (SS-LGF-OFS) and the KNN classifier; S-scale, O-Orientation. 98

Figure 5-13 APIA% for stress classification using SS-ALGF-OFS and SS-SP-ALGF-OFS features. 99

Figure 5-14 APIA% for emotion classification for each of the 12 log-Gabor filters (SS-LGF-OFS) using the GMM and KNN classifiers; S-scale, O-orientation. 100

Figure 5-15 APIA% for emotion classification using SS-ALGF-OFS and SS-SP-ALGF-OFS features. 101

Figure 5-16 APIA (%) versus filter orientation (degree) for two scales m=1 (blue line) and m=2 (pink line) for words with only ā vowel and the GMM classifier. 102

Figure 5-17 Framework of stress and emotion recognition based on sigma-pi unit. 103

Figure 5-18 Structure of a Sigma-pi neuron. 105

Figure 5-19 Generation of Si values over speech frames. 105

Figure 5-20 APIA % for stress classification using the SS-sigma-pi features (48ms time distance) within CB, Bark and ERB bands and the GMM and KNN classifier. 107

Figure 5-21 APIA % for emotion classification using the SS-sigma-pi features (48ms time distance) within CB, Bark and ERB bands and the GMM and KNN classifiers. 108

Figure 5-22 The WP-ALGF-OFS feature extraction process. 110

Figure 5-23 Wavelet Packet (WP) two-levels decomposition; Gi denotes the low pass filters, Hi denotes the high pass filters, ↓2 denote dyadic division by a factor of 2. 110

Figure 5-24 Examples of wavelet packet arrays for the vowel ō in the word “break” pronounced under (a) high-level stress speech, (b) low-level stress speech and (c) neutral speech. Horizontal axis-time, vertical axis-frequency; the color indicates the coefficient’s values: purple-high values of coefficients, blue-low values of coefficients. 111

Figure 5-25 Examples of wavelet packet arrays for utterances pronounced with the following emotions: (a) angry, (b) anxious, (c) dysphoric, (d) neutral and (e) happy. Note that each emotion is represented by a different utterance. Horizontal axis-time, vertical axis-frequency; the color indicates the coefficient’s values: purple-high values of coefficients, blue-low values of coefficients. 112
Figure 5-26 APIA% for stress classification using WP-ALGF-OFS features and two classifiers GMM and KNN. ................................................................. 114
Figure 5-27 A comparison of APIA% for stress and emotion classification using SS-ALGF-OFS, SS-ERB-AE and WP-ALGF-OFS features. ...................... 116
Figure 5-28 Feature extraction using EMD and Renyi entropy (ER). ......................... 117
Figure 5-29 APIA% for stress classification using EMD-AER features with the Renyi’s entropy of order q=2 and q=3. ......................................................... 119
Figure 5-30 APIA% for emotion classification using EMD-AER features with the Renyi’s entropy of order q=2 and q=3. ......................................................... 120
Figure 6- 1 New nonlinear model of the glottal flow formation based on [91] and [75, 223]. ................................................................................................................. 124
Figure 6- 2 An example feature using the area under TEO autocorrelation contour. .... 127
Figure 6- 3 Flowchart of the TEO based feature extraction process. ......................... 127
Figure 6- 4 Frequency responses of the perceptual wavelet packet (PWP) decomposition filters (using db2 wavelet). ....................................................... 130
Figure 6- 5 APIA% for stress classification using TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G features combined with the GMM classifier. .... 131
Figure 6- 6 APIA% for stress classification using TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G features combined with the KNN classifier. .... 131
Figure 6- 7 APIA% for emotion classification using TEO-CB, TEO-DWT, TEO-WP, TEO-PWP features and the GMM and KNN classifiers ......................... 133
Figure 6- 8 APIA% of stress classification for TEO-PWP-S features combined with MLPNN and PNN classifiers ............................................................ 136
Figure 6- 9 Calculation of the Teager Energy Operator contour values. ................. 137
Figure 6- 10 APIA% for stress classification using TEO-contour based feature sets combined with the GMM classifier. ............................................. 140
Figure 6- 11 APIA% for stress classification using TEO-contour based feature sets combined with the KNN classifier. ............................................. 140
Figure 6- 12 APIA% for emotion classification using TEO-contour based feature sets combined with the GMM and KNN classifiers. .......................... 141
Figure 6- 13 Examples of the spectral energies of the glottal waveforms for the word “break” under different levels of stress pronounced by the same speaker. .... 143
Figure 6- 14 Examples of the spectral energies of the glottal waveforms produced with different emotions. ................................................................. 143
Figure 6- 15 Calculation steps for AUSEEG and AUSEES features. .......................... 144
Figure 6- 16 APIA% for stress classification using AUSEEG and AUSEES features. .. 146
Figure 6- 17 APIA% for emotion classification using AUSEEG and AUSEES features. ................................................................. 147
Figure 7- 1 General flowchart of the emotion classification system. ...................... 153
Figure 7- 2 Emotion classification accuracy (APIA %) for parents of depressed adolescents (PD) and parents of non-depressed adolescents (PND). ......... 155
Figure 7- 3 Emotion classification accuracy (APIA %) for mothers and fathers. ........ 156
Figure 7- 4 Correct recognition rates for each emotion for parents of depressed (PD) and non-depressed adolescents (PND) using TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG features and the GMM classifier; Co-contempt, Ag-angry, Ax-anxious, Dy-dysphoric, Pl-pleasant, Ne-neutral, Ha-happy. ................................. 157
Figure 7- 5 Gender based classification accuracy for each emotion using TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG features and the GMM classifier; Co-contempt, Ag-angry, Ax-anxious, Dy-dysphoric, Pl-pleasant, Ne-neutral, Ha-happy.
................................................................................................................................. 157

Figure 8- 1 APIA % for all features using the GMM classifier................................. 169
Figure 8- 2 APIA % for all features using the KNN classifier................................. 169
Figure 8- 3 An average percentage of the energy contribution by each of the 16 linear bands to the total energy within a frame of the glottal wave for different stress levels. These results were obtained using the AUSEEG features and the GMM classifier................................................................. 171
Figure 8- 4 An average percentage of the energy contribution by each of the 16 linear bands to the total energy within a frame of the glottal wave for different emotions. These results were obtained using the AUSEEG features and the GMM classifier. ................................................................. 171
Figure 8- 5 Emotion classification (APIA %) using AUSEEG-LB/GMM within selected band................................................................. 172
Figure 8- 6 Emotion classification using AUSEEG/GMM within selected frequency bands. ................................................................. 176
**List of Acronyms**

- **APIA**: Average Percentage of Identification Accuracy
- **AUSEEG**: the area under the glottal wave energy envelope calculated within spectral sub-bands
- **AUSEES**: the area under the speech wave energy envelope calculated within spectral sub-bands
- **DWT**: Discrete Wavelet Transform
- **EMD**: Empirical Mode Decomposition
- **EMD-AER**: features derived from the empirical mode decomposition combined with calculation of an averaged Renyi’s entropy
- **ERB**: Equivalent Rectangular Bandwidth
- **MFCC**: Mel-Frequency Cepstral Coefficient
- **MI**: Mutual Information
- **MLPNN**: Multilayer Perceptron Neural Network
- **ORI**: Oregon Research Institute
- **PNN**: Probabilistic Neural Network
- **SS-AF-CB-AE**: features extracted from speech spectrograms with anisotropic filtering and critical band scales
- **SS-AF-BARK-AE**: features extracted from speech spectrograms with anisotropic filtering and Bark scales
- **SS-AF-ERB-AE**: features extracted from speech spectrograms with anisotropic filtering and equivalent rectangular bandwidth scales
- **SS-CB-AE**: features extracted from speech sub-bands of spectrograms using critical band scales
- **SS-BARK-AE**: features extracted from speech sub-bands of spectrograms using Bark scales
- **SS-ERB-AE**: features extracted from speech sub-bands of spectrograms using equivalent rectangular bandwidth scales
- **SS-ALGF-OFS**: features extracted from speech spectrograms and combined with a log-Gabor filter-bank, averaging and an optimal feature selection
- **SS-LGF-OFS**: features extracted from speech spectrograms and combined with a single log-Gabor filter and an optimal feature selection
- **SS-SP-ALGF-OFS**: feature generation using spectrogram patches and log-Gabor filters
<table>
<thead>
<tr>
<th>SS-sigma-pi</th>
<th>features extracted from speech spectrograms using sigma-pi units</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUSAS</td>
<td>Speech Under Simulated and Actual Stress</td>
</tr>
<tr>
<td>TEO</td>
<td>Teager Energy Operator</td>
</tr>
<tr>
<td>TEO-CB</td>
<td>the area under the normalized TEO autocorrelation envelope calculated within critical bands</td>
</tr>
<tr>
<td>TEO-DWT</td>
<td>the area under the normalized TEO autocorrelation envelope calculated within octave bands of the discrete wavelet transform decomposition</td>
</tr>
<tr>
<td>TEO-WP</td>
<td>the area under the normalized TEO autocorrelation envelope calculated within octave bands of the wavelet packet decomposition</td>
</tr>
<tr>
<td>TEO-PWP</td>
<td>the area under the normalized TEO autocorrelation envelope calculated within octave bands of the perceptual wavelet packet decomposition</td>
</tr>
<tr>
<td>WP-ALGF-OFS</td>
<td>features extracted from the wavelet packet arrays combined with a log-Gabor filter bank, averaging and an optimal feature selection</td>
</tr>
</tbody>
</table>
Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; and, any editorial work, paid or unpaid, carried out by a third party is acknowledged.

Ling He

November 2010
Acknowledgements

I would like to express my deep and sincere gratitude to my supervisor, Dr Margaret Lech. Her knowledge, support, enthusiasm and the academic guidance have been of great value to me.

I would also like to thank Professor Nicholas Allen from the Department of Psychological Sciences, University of Melbourne and Dr Lisa Sheeber from the Oregon Research Institute, U.S.A. for providing a data base of emotional speech. Their kind advice was most helpful during my candidature.

My sincere thanks also go to Dr Namunu C. Maddage for his much appreciated advice and to my colleagues from the DSP Lab, RMIT University: Lu-Shih Alex Low and Sheeraz Memon.

I would also like to thank my advisor during my Master’s study, Professor Hua Huang, from the Department of Telecommunication, Sichuan University, who introduced me to the filed of signal processing.

Special thanks are due to my parents Xianpei He and Suzhen Deng. They worked very hard to support the family and to raise me up. Thanks for their unconditional love and support throughout my life.

Last but not least, many thanks to my husband Zhengbo Zhang. Without his encouragement and understanding, it would have been impossible for me to finish my doctoral study.
Abstract

The speech stress and emotion recognition and classification technology has a potential to provide significant benefits to the national and international industry and society in general. The accuracy of an automatic emotion speech and emotion recognition relays heavily on the discrimination power of the characteristic features.

This work introduced and examined a number of new linear and nonlinear feature extraction methods for an automatic detection of stress and emotion in speech.

The proposed linear feature extraction methods included features derived from the speech spectrograms (SS-CB/BARK/ERB-AE, SS-AF-CB/BARK/ERB-AE, SS-LGF-OFS, SS-ALGF-OFS, SS-SP-ALGF-OFS and SS-sigma-pi), wavelet packets (WP-ALGF-OFS) and the empirical mode decomposition (EMD-AER).

The proposed nonlinear feature extraction methods were based on the results of recent laryngological studies and nonlinear modelling of the phonation process. The proposed nonlinear features included the area under the TEO autocorrelation envelope based on different spectral decompositions (TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G), as well as features representing spectral energy distribution of speech (AUSEES) and glottal waveform (AUSEEG).

The proposed features were compared with features based on the classical linear model of speech production including F0, formants, MFCC and glottal time/frequency parameters.

Two classifiers GMM and KNN were tested for consistency. The experiments used speech under actual stress from the SUSAS database (7 speakers; 3 female and 4 male) and speech with five naturally expressed emotions (neutral, anger, anxious, dysphoric and happy) from the ORI corpora (71 speakers; 27 female and 44 male).
The nonlinear features clearly outperformed all the linear features. The classification results demonstrated consistency with the nonlinear model of the phonation process indicating that the harmonic structure and the spectral distribution of the glottal energy provide the most important cues for stress and emotion recognition in speech.

The study also investigated if the automatic emotion recognition can determine differences in emotion expression between parents of depressed adolescents and parents of non-depressed adolescents. It was also investigated if there are differences in emotion expression between mothers and fathers in general. The experiment results indicated that parents of depressed adolescent produce stronger more exaggerated expressions of affect than parents of non-depressed children. And females in general provide easier to discriminate (more exaggerated) expressions of affect than males.


**Patents and Publications**

**Patents**

1. Provisional Patent Application: 2010902987

**Book Chapters (peer reviewed):**


**Journal Articles (peer reviewed):**


**Conference Proceedings (peer reviewed):**


Chapter 1. Introduction

This Chapter provides the problem statement, thesis aims and the thesis original contributions. An outline of the thesis structure is also given.

1.1 Problem statements

In recent years, speech scientists and engineers, who had tended to disregard pragmatic and paralinguistic aspects of speech in their effort to develop models of speech communication for speech technology applications, have started to devote more attention to speaker attitudes and emotions—often in the interest to increase the acceptability of speech technology for human users [16, 17].

The speech signal communicates linguistic information between speakers as well as paralinguistic information about speaker’s emotions, personalities, attitudes, feelings, levels of stress and current mental states. Just as effective human-to-human communication is virtually impossible without speakers being able to detect and understand each other’s emotions, human-machine communication suffers from significant inefficiencies because machines cannot understand our emotions or generate emotional responses. Words are not enough to correctly understand the mood and intention of a speaker and thus the introduction of human social skills to human-machine communication is of paramount importance. This can be achieved by the researching and creating methods of speech modeling and analysis that embrace the signal, linguistic and emotional aspects of communication.

Speech stress and emotion recognition and classification is aimed to automatically detect stress and emotions in speech signals through analyzing the vocal behavior as a marker of affect (e.g. emotions, moods and stresses), focusing on the nonverbal aspect of speech. The assumption is that most affective states involve the physiological reactions, like the changes in the autonomic and somatic nervous systems, which in turn influences
the speech production process, and causes the changes of several voice parameters [16, 17].

The automatic affect recognition in speech has applications in behavioral and mental health science, human to machine communication, social robotics, security systems, computer games, online education, commercial field and medicine. It can be used to improve the robustness of speech and speaker recognition systems [61-63]. Moreover, by assessing a speaker’s stress level, stress classification can support automatic assessment of mental states of people working in dangerous environments (e.g. chemicals, explosives) and people undertaking high level of responsibility (e.g. pilots, surgeons). Various clinical applications of affect analysis in speech have been reported in diagnosis of conditions such as depression [30, 81-85], autism [86], Alzheimer and dementia [88], and schizophrenia [89]. These works demonstrate importance of speech as a diagnostic signal providing valid information about speaker’s mental and physiological state. Call systems can use emotion recognition to sort emergency telephone messages, or cope with dispute through monitoring the mental states (levels of satisfaction) of customers. Another commercial application of emotion detection system is the interactive game industry offering the sensation of naturalistic human-like interaction, and the entertainment industry facilitating the intelligent agents with sensitivity to player’s mood as well as the ability to response accordingly through affective voice or face expression.

1.2 Thesis aims

The proposed research project aimed to:

- develop and test new efficient feature extraction methods for an automatic recognition and classification of stress and emotion in speech;

- determine if the recent laryngological studies indicating a nonlinear character of speech production can be used to derive efficient feature extraction methods for an automatic recognition and classification of stress and emotion in speech.
1.3 Thesis scope

The research was focused only on the natural speech i.e. speech with naturally expressed (not acted) stress and emotions. Two databases containing natural speech were used in this project:

- The Speech under Stress and Actual Stress database comprised of speech samples representing a wide variety of acted and actual stresses and emotions recorded from 32 speakers (13 male and 19 female). The speech was recorded within work environment (pilot cabin) and during a rollercoaster ride. The speakers were asked to read a set of 35 words listed in a random order.

- The Oregon Research Institute (ORI) database recorded by the psychologists for the purpose of behavioral studies contained recordings from 170 speakers (75 male and 95 female). The recordings were made during family conversation.

In the automatic classification of stress 3 different levels of stress were classified: high level stress, low level stress and neutral speech. The automatic emotion classification study tested up to 7 different emotions: contempt, angry, anxious, dysphoric, pleasant, neutral and happy.

All speakers had general-USA English accent. The emotion classification tests had text-, speech-, and speaker-independent character. In most cases with exception in Chapter 7, the tests were also gender-independent. The speech bandwidth tested in the experiments was 4 kHz (Chapters 4, 5 and 6) and 11 kHz (Chapter 7).

1.4 Thesis original contributions

This thesis proposed and tested several new feature extraction methods for stress and emotion detection, the results were compared with existing classical approaches.

The proposed new feature extraction methods include the following linear and nonlinear methods.
Linear methods:
- parameters extracted from speech sub-bands of spectrograms without anisotropic filtering (SS-CB-AE, SS-BARK-AE, SS-ERB-AE),
- parameters extracted from sub-bands of speech spectrograms with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE, SS-AF-ERB-AE),
- parameters extracted from speech spectrograms and combined with a single log-Gabor filter and an optimal feature selection (SS-LGF-OFS),
- parameters extracted from speech spectrograms and combined with a log-Gabor filter-bank, averaging and an optimal feature selection (SS-ALGF-OFS),
- parameters extracted from speech spectrograms using spectrogram patches and log-Gabor filters (SS-SP-ALGF-OFS),
- parameters extracted from speech spectrograms and sigma-pi units (SS-sigma-pi),
- parameters extracted from the wavelet packet arrays combined with a log-Gabor filter-bank, averaging and an optimal feature selection (WP-ALGF-OFS),
- parameters derived from the Empirical Mode Decomposition (EMD) combined with calculation of an averaged Renyi’s entropy (EMD-AER).

Nonlinear methods:
- features derived from the Teager Energy Operator (TEO) applied to the speech signal (TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S),
- features derived from the Teager Energy Operator (TEO) applied to the glottal waveform (TEO-PWP-G),
- features derived from the spectral energy of speech (AUSEES),
- features derived from the spectral energy of the glottal wave (AUseeG).

1.5  Major findings and conclusions

The study led to the following major findings and conclusions:

- The nonlinear TEO based features (TEO-CB, TEO-WP and TEO-PWP), as well as AUSEES and AUseeG features showed very close performance and clearly outperformed all other features.
− It was found that the TEO analysis provided the best performance when applied to spectral sub-bands based on the human auditory perception (critical bands). Other types of features such as the spectrogram based features also showed improved performance when combined with spectral subdivision based on an auditory scale.

− The outstanding performance of the nonlinear features based on recent laryngological experiments with the acoustic flow formation pointed to the importance of nonlinear mechanisms associated with the glottal flow formation as important cues for stress and emotion.

− It was demonstrated that automatic emotion classification can be used to test some of the psychological hypotheses. It was showed that mothers in general express their emotions in a stronger way than fathers, and the parents of depressed adolescents in general express their emotions in a stronger way than the parents of non-depressed adolescents, which was consistent with previous psychological observations [229, 230, 232].

1.6 Thesis outline

The thesis is organized as follows:

Chapter 1 provides the problem statement, thesis aims and the thesis original contributions. An outline of the thesis structure is also given.

Chapter 2 introduces definitions of stress and emotion. It describes types of emotions and emotional class labeling. The importance of the research on emotional speech analysis is explained. The Chapter also provides a review of emotional speech data collections and the existing methods of feature extraction, and emotion modeling techniques.
Chapter 3 describes contents of the speech corpora used in this study. It introduces a general framework of stress and emotion detection and describes methods used in the pre-processing, feature extraction and classification stages of this framework.

Chapter 4 describes stress and emotion classification experiments based on classical features including: fundamental frequency $F_0$, formants ($F_1$, $F_2$ and $F_3$), mel frequency cepstral coefficients (MFCC) and glottal features. The classical features represent parameters describing a classical source-filter model of speech production.

Chapter 5 introduces a number of new feature extraction methods for an automatic detection of stress and emotion in speech. The proposed feature extraction methods use features derived from the speech spectrograms (SS-CB/BARK/ERB-AE, SS-AF-CB/BARK/ERB-AE, SS-LGF-OFS, SS-ALGF-OFS, SS-SP-ALGF-OFS, SS-sigma-pi), wavelet packets (WP-ALGF-OFS) and the empirical mode decomposition (EMD-AER). The proposed features are tested using the same speech data sets and classifiers as those used with classical features tested in Chapter 4.

Chapter 6 introduces a number of new feature extraction methods for an automatic detection of stress and emotion in speech (TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S/G, AUSSEG, AUSSES). The proposed feature extraction methods were inspired by a number of recent laryngological experiments and new nonlinear models of speech production. The proposed features were tested using the same speech data sets and classifiers as those used with classical features tested in Chapter 4.

Chapter 7 investigates an automatic emotion classification in spontaneous speech within two family environments: a family with parents and adolescents diagnosed as clinically depressed and a family with parents and adolescent who are non-depressed. The emotion classification results are also used to determine if the classification rates differ between speakers who are parents of depressed adolescents and speakers who are parents of non-depressed adolescents. The study also investigated the effect of gender on emotion classification by looking at the classification rates obtained for mothers and fathers. The classification experiments used the best performing feature extraction
methods (TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG) determined in Chapter 6 and two classifiers: GMM and KNN. The speech data used in this chapter was extracted from the ORI data base and extended to include not 5 but 7 different emotions: contempt, angry, anxious, dysphoric, pleasant, neutral and happy. The speech bandwidth was extended from 4 kHz to 11 kHz.

Chapter 8 compares a number of features based on the nonlinear model of speech production proposed in Chapter 7 with some of the classical features based on the linear model and described in Chapter 6. The nonlinear features include the TEO based features (TEO-CB/DWT/WD/PWP), as well as features representing spectral energy distribution of speech (AUSEES) and glottal waveform (AUSEEG). These features are compared the classical features including: the fundamental frequency F0, formants and MFCC. Two classifiers GMM and KNN were tested for consistency. The experiments used speech under actual stress from the SUSAS database (7 speakers; 3 female and 4 male) and speech with five naturally expressed emotions (neutral, anger, anxious, dysphoric, and happy) from the ORI corpora (71 speakers; 27 female and 44 male). The classification results demonstrated consistency with the nonlinear model of the phonation process indicating that the harmonic structure and the spectral distribution of the glottal energy provide the most important cues for stress and emotion recognition in speech. The frequency range from 2.5 kHz to 4 kHz provided the most discriminative features which were on their own almost as effective in class discrimination as features representing the entire speech bandwidth (4 kHz).

Chapter 9 summarizes the results of this thesis and discusses some of the problems connected with the findings. A performance based ranking list of all stress and emotion classification methods researched in this thesis is given.
Chapter 2. Review of Existing Methods of Stress and Emotion Classification in Speech

This Chapter introduces definitions of stress and emotion. It describes types of emotions and emotional class labeling. The importance of the research on emotional speech analysis is explained. The Chapter also provides a review of emotional speech data collections and the existing methods of feature extraction, and emotion modeling techniques.

2.1 What are stress and emotions?

This section introduces definitions of stress and emotion and ways of describing different emotional states.

2.1.1 Definitions of stress and emotion

Both stress and emotion are psycho-physiological states involving characteristic somatic and autonomic responses [1].

Stress is a psychological and biological term characterized by loss of ability to appropriately respond to difficult emotional and physical conditions that can be either real or imagined. Stress is characterized by subjective strain, dysfunctional physiological activity, and deterioration of performance [2]. Typical stress symptoms include a state of high arousal, increased heart rate, excessive adrenaline production, failure of the coping mechanisms, feeling of strain and exhaustion and inability to concentrate. Stress may be induced by external factors (workload, noise, vibration, sleep loss, etc.) or by internal factors (fatigue, etc.) [3].

Existing stress detection and classification research uses stress categories (or classes) bases on different levels of difficulties that a given person has to deal with [4].
Emotion comprises complex psychological and physiological phenomena including person’s state of mind and the way an individual interacts with the environment. Generally, emotion involves: physiological arousal, expressive behaviors, and conscious experience [5].

Emotion is closely related to the state called mood. Unlike emotion which a short-term (minutes to hours) psycho-physiological state, mood is a relatively long lasting emotional state (hours-weeks). In contrast to simple emotions, moods are less specific, less intense, and less likely to be triggered by a particular stimulus or event [6].

2.1.2 Types of emotions

One of the major problems in the area of emotion detection and classification is the lack of consistent definitions of emotions and a lack of a unified qualitative system dividing emotions into different categories.

Studies by Bradley [7], Lang et al.[8], Osgood et al. [9], Russel and Mehrabian [10], introduced 3-dimensional approach to the emotion description. The 3 dimensions of emotion include:

- Valence
- Arousal
- Control (Dominance)

Valence represents types of emotions which on a continuous scale extend from pleasant or “positive” to unpleasant or “negative” [11], while arousal is characterized by the intensity of emotional states ranging on a continuous scale from energized, excited and alert to calm, drowsy or peaceful. The third dimension control (dominance) is used to distinguishing among emotional states having similar arousal and valence and ranges from “no control” to “full control”.

Based on the above measures, human emotions can be represented as points in the 3-dimensional space [12] illustrated in Figure 2-1. Since the “control” variable in Figure 2-
1 occupies a relatively narrow range, in many cases it is disregarded and emotions are described in a 2-dimensional arousal versus valence space.

![Emotion Space Diagram](image)

**Figure 2-1** Areas and bounders of the 3-D emotion space (adapted from Dietz and Lang [13]).

Although, there is a steady increasing body of research using the continuous 3-dimensional (arousal, violence and control) or 2-dimensional (arousal and violence) description of emotional states [12], the majority of studies concerned with an automatic detection and classification of emotion in speech use a categorical approach to the discrimination between different emotional states.

The categorical approach is particularly useful in the context of the pattern recognition methodology since it divides emotions into a finite number of categories (or classes). A large number of emotional categories have been proposed in literature and a comprehensive list of different emotional labels and their descriptions can be found in Table 2-1 which is based on Whissell [14], Plutchik [15] and Cowie [16].

One of the major drawbacks of the categorical approach is a lack of a unified system for emotional class labelling. Existing class labels vary between different data bases. This problem is not easy to solve mostly due to subjective nature in which emotions are perceived. Differences in individual perception can lead to very different labelling of the same data by different assessors. The consistency of a subjective labelling of emotions is
usually restricted to databases using the same ways of subjective assessments and the same assessors.

| Table 2-1 Emotion Labels in alphabetical ordering (based on Whissell [14], Plutchik [15] and Cowie [16]). |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Accepting | Bashful | Envious | Co-operative | Planful | Self-controlled | Sad |
| Adventurous | Bewildered | Exasperated | Critical | Pleased | Serene | Sarcastic |
| Affectionate | Bitter | Expectant | Curious | Possessive | Shy | Satisfied |
| Afraid | Boastful | Forlorn | Daring | Hopeless | Sociable | Scared |
| Aggressive | Bored | Furious | Defiant | Hostile | Sorrowful | Scornful |
| Agreeable | Calm | Generous | Delighted | Humiliated | Stubborn | Timid |
| Amazed | Cautious | Gleeful | Demanding | Impatient | Submissive | Tolerant |
| Ambivalent | Cheerful | Gloomy | Depressed | Impulsive | Surprised | Trusting |
| Amused | Disgusted | Greedy | Despairing | Indecisive | Suspicious | Unaffectionate |
| Angry | Disinterested | Grief-stricken | Disagreeable | Indignant | Sympathetic | Uncertain |
| Annoyed | Disobedient | Grouchy | Disappointed | Inquisitive | Puzzled | Uncooperative |
| Antagonistic | Displeased | Guilty | Discouraged | Interested | Quarrelsome | Unfriendly |
| Anticipatory | Dissatisfied | Happy | Nervous | Intolerant | Ready | Unhappy |
| Anxious | Distrustful | Helpless | Obedient | Irritated | Receptive | Unreceptive |
| Apathetic | Eager | Hesitant | Obliging | Jealous | Reckless | Unsympathetic |
| Apprehensive | Ecstatic | Hopeful | Outraged | Joyful | Rebellious | Vascillating |
| Ashamed | Elated | Confused | Panicky | Loathful | Rejected | Vengeful |
| Astonished | Embarrassed | Contemptuous | Patient | Lonely | Remorseful | Watchful |
| Attentive | Empty | Content | Pensive | Meek | Resentful | Wondering |
| Awed | Enthusiastic | Contrary | Perplexed | Self-conscious | Revolted | Worried |

In reality, it is not possible to achieve automatic emotion recognition with resolution that would allow distinguishing between all of the emotional labels listed in Table 2-1. The current recognition systems can provide reasonably good differentiation between 2-8 types of basic emotions. The most often used emotional class labels are listed in Table 2-2.
In this thesis speech data representing three levels of stress: high stress, low stress and neutral, and seven different types of emotions: contempt, angry, anxious, dysphoric, pleasant, happy and neutral were used. The details were described in Chapter 3.1.

2.1.3 Simulated (acted), elicited and natural emotions

The speech based communication of naturally expressed emotion is based on the fact that different types of emotion are characterized by unique patterns of acoustic cues [17]. A lack of such patterns would make differentiation between emotions impossible.

The major question arising when emotions are artificially elicited in speakers is whether these unique patterns of acoustic cues are still produced.

Based on the way the emotions were evoked in speakers, the existing speech data can be classified into three major categories [17]:

(1) Natural vocal expression

Speech data containing speech during naturally occurring emotional states has been collected in the past during dangerous flight situations for pilots, journalists reporting emotion eliciting events, affectively loaded therapy sessions, or talk and game shows on TV [17] [18]. The use of natural emotions represents an ideal situation from the research point of view, since it provides the most valid and naturally occurring emotional cues. Therefore the ecological validity of such data is very high. However, there are some serious problems with controlling and designing data collection procedures that can provide a sufficient amount and variety of truly natural emotional speech. It is extremely difficult to collect data with the same intensity and types of naturally expressed emotions from a large number of speakers.

(2) Induced emotional expression

Emotional speech has been collected from speakers after specific emotional states were artificially induced in groups of speakers. The emotional states were induced using
psychoactive drugs. For example, Helfrich et al. [19] observed the effects of antidepressive drugs on vocal parameters and compared to placebo. The majority of methods use stress induction by arranging difficult tasks to be completed under time pressure, the presentation of emotion inducing films or slides, or imagery methods [2] [20]. The approach based on induced emotions provides experimental psychologists with high degree of control allowing for generation of comparable speech samples for all participants. The major drawbacks include the fact that the emotion induction procedures often produce only relatively weak emotions. In spite of using the same procedures for all participants, it is not certain that similar emotional states are produced in all speakers.

(3) Simulated (acted) emotional expression

This is the most frequently used approach to provide emotional speech samples. Professional actors are usually asked to produce vocal expressions of emotion (often using standard verbal content) and typical labels of emotional categories [21, 22]. This approach yields much more intense and highly controlled, emotional expressions compare to the induced or natural states. The disadvantage is that it cannot be excluded that actors overemphasize relatively obvious cues and miss more subtle ones that might appear in natural expression of emotion. In [17] it is argued that acted emotional expressions reflect sociocultural norms rather than the psychophysiological effects on the voice.

The speech data used in this thesis contained only naturally expressed stress and emotions.

2.1.4 Data labeling (annotation) methods

The speech annotation process which follows the collection of emotional speech data aims to provide piece-wise emotional labelling or annotation of the speech recordings. The labelling process based on a subjective assessment of emotions expressed through speech is usually done by:
(1) **Self-assessment made by the speakers**

After the speech recordings are made, speakers can be asked to listen to these recordings and describe emotions that they were expressing.

(2) **Assessment made by independent listeners**

A number of randomly chosen listeners without prior training can be asked to listen to speech recordings and provide emotional labelling usually by choosing the most suitable label from a provided list of possible emotions. Audiovisual data bases collected for the purposes of psychological and behavioural research use annotation systems with second-by-second annotation provided by highly trained psychologist. This annotation is usually based on a large range of acoustic and visual cues.

(3) **Assessment of elicited emotions**

When emotions are elicited either by drugs or arrangement of situations evoking certain types of emotions, it is assumed that the appropriate emotions are expressed. In some cases, the correctness of elicited emotions needs to be confirmed by self-assessment or assessment made by independent listeners.

It has to be noted that an independent assessment by listeners in principle does not need to agree with the emotions actually felt by the speakers; therefore the choice between self-assessment and independent assessment depends on the purpose for which these assessments are needed. For example, an independent assessment by a group of listeners is often used to make a comparison between machine based and human based emotion recognition in speech. In both cases, the assessment of performance is based on the annotation representing the emotions actually felt by speakers.

The data used in this thesis contained speech under elicited stress and speech data with naturally expressed emotions annotated by highly trained psychologists.
2.1.5 Existing emotional speech data collections

The classification accuracy of stress and emotion depends largely on the type of speech samples used in the process of statistical modeling of different classes of stress or emotion. Previous research used three types of speech data collections. The first type used emotions simulated by professional actors in a recording laboratory allowing experimental control but having low ecological validity. The second type of data represented natural vocal expressions recorded in the field or from reality media broadcasts. It provided high ecological validity but it was difficult to determine the actual emotion felt by the speaker. The third type used experimentally induced emotional expressions in the laboratory. This approach provided low level of control over the emotional arousal and valence.

Apart from ecological validity, other important factors characterizing the quality of emotional speech data are the size of the speech data base (number of speech samples) and the accuracy of emotional labeling (or data annotation).

Too small sets will not provide sufficient number of voice cues for training and testing of automatic stress and emotion recognition systems. Unfortunately the optimal choice of data has to be often compromised by restrictions in data availability, as well as ethical and practical difficulties in conducting experiments which provide highly natural expressions of emotions.

Table 2-2 contains a list of several widely used emotional databases described in terms of numbers of subject, language, types of emotions, and naturalness of the speech (e.g. simulated, natural or semi-natural) [23].

Most of the databases in Table 2-2 were collected in a lab environment for the purpose of emotion recognition studies. The number of subjects was restricted to a small range of 1 to 238 subjects and therefore in some cases the number of speech utterances was also relatively small compare for example with the NIST speaker recognition corpora which
usually contain a few hundreds of speaker recorded over a number of sessions spread over a few months period.

**Table 2-2** Description of emotional databases.

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of subjects</th>
<th>Type of emotions</th>
<th>Language</th>
<th>Naturalness</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Danish Emotional Speech database [24]</td>
<td>4</td>
<td>Anger, happiness, neutrality, sadness and surprise</td>
<td>Danish</td>
<td>Simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>ELRA corpus number S0020 [25]</td>
<td>238</td>
<td>Only partially oriented to emotion</td>
<td>Dutch</td>
<td>Simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>Berlin Emotions Databases [26]</td>
<td>10 (5 male and 5 female)</td>
<td>Anger-hot, boredom, disgust, fear-panic, happiness, neutrality, sadness-sorrow</td>
<td>German</td>
<td>Simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>van Bezooijen [27]</td>
<td>8 (4 male and 4 female)</td>
<td>Anger, contempt, disgust, fear, interest, joy, neutrality, sadness, shame, surprise</td>
<td>Dutch</td>
<td>Simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>Belfast structured database [28]</td>
<td>50</td>
<td>Anger, fear, happiness, neutrality, sadness</td>
<td>English</td>
<td>Simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>Belfast natural database [28]</td>
<td>125 (31 male and 94 female)</td>
<td>Wide range (active positive emotion, active negative emotion, passive positive emotion, passive negative emotion)</td>
<td>English</td>
<td>Natural</td>
<td>Audio-visual</td>
</tr>
<tr>
<td>ORI database [29]</td>
<td>170 (75 male and 95 female)</td>
<td>Contempt, anger, anxiety, dysphoria, pleasantness, neutrality and happiness</td>
<td>English</td>
<td>Natural</td>
<td>Audio-visual</td>
</tr>
<tr>
<td>Abelin et al. [31]</td>
<td>1</td>
<td>Anger, disgust, dominance, fear, joy, sadness, shyness, surprise</td>
<td>Swedish</td>
<td>Simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>Banse and Scherer [32]</td>
<td>12 (6 male and 6 female)</td>
<td>Anger (hot), anger (cold), anxiety, boredom, contempt, disgust, elation, fear (panic), happiness, interest, pride, sadness, shame</td>
<td>German</td>
<td>Semi-natural</td>
<td>Audio-visual</td>
</tr>
<tr>
<td>Polzin et al. [33]</td>
<td>unspecified</td>
<td>Anger, sadness, neutrality</td>
<td>English</td>
<td>Simulated</td>
<td>Audio-visual</td>
</tr>
<tr>
<td>Chung [34]</td>
<td>6</td>
<td>Joy, neutrality, sadness</td>
<td>English and Korean</td>
<td>Natural</td>
<td>Audio-visual</td>
</tr>
<tr>
<td>Amir et al. [35]</td>
<td>61</td>
<td>Anger, disgust, fear, joy, neutrality, sadness</td>
<td>Hebrew and Russian</td>
<td>Simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>Iriondo et al. [36]</td>
<td>8</td>
<td>Desire, disgust, fury, fear, joy, surprise, sadness</td>
<td>Spanish</td>
<td>Semi-natural</td>
<td>Audio</td>
</tr>
<tr>
<td>Alter et al. [37]</td>
<td>1</td>
<td>Anger (cold), happiness, neutrality</td>
<td>German</td>
<td>Simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>Yu et al. [38]</td>
<td>unspecified</td>
<td>Anger, fear, joy, neutrality, sadness</td>
<td>Chinese</td>
<td>Simulated</td>
<td>Audio-visual</td>
</tr>
<tr>
<td>Yildirim et al. [40]</td>
<td>1</td>
<td>Anger, happy, neutrality, sadness</td>
<td>English</td>
<td>Simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>Liberman [41]</td>
<td>unspecified</td>
<td>Wide range</td>
<td>English</td>
<td>Simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>Petrushin [42]</td>
<td>30</td>
<td>Anger, fear, happiness, neutrality, sadness</td>
<td>English</td>
<td>Simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>Clavel et al. [43]</td>
<td>18</td>
<td>Neutrality, levels of fear</td>
<td>English</td>
<td>Simulated</td>
<td>Audio-visual</td>
</tr>
<tr>
<td>Fischer [44]</td>
<td>56</td>
<td>Anger, depression, neutrality</td>
<td>German</td>
<td>Natural</td>
<td>Audio</td>
</tr>
<tr>
<td>Lee, Narayanan [45]</td>
<td>unspecified</td>
<td>Negative-positive</td>
<td>English</td>
<td>Natural</td>
<td>Audio</td>
</tr>
<tr>
<td>Linnankoski et al. [46]</td>
<td>13</td>
<td>Wide range</td>
<td>English</td>
<td>Semi-natural</td>
<td>Audio</td>
</tr>
<tr>
<td>Lloyd [47]</td>
<td>1</td>
<td>Phonological stress</td>
<td>English</td>
<td>simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>Martins et al. [48]</td>
<td>10</td>
<td>Anger, disgust, happiness, irony</td>
<td>Portuguese</td>
<td>simulated</td>
<td>Audio</td>
</tr>
<tr>
<td>McMahon et al. [49]</td>
<td>29</td>
<td>Annoyance, shock, stress</td>
<td>English</td>
<td>Semi-natural</td>
<td>Audio</td>
</tr>
<tr>
<td>Polzin and Waibel [50]</td>
<td>5</td>
<td>Anger, fear, happiness, neutrality, sadness</td>
<td>English</td>
<td>simulated</td>
<td>Audio</td>
</tr>
</tbody>
</table>
The majority of emotional databases in Table 2-2 contain only simulated or semi-natural expressions of emotions. The natural emotional speech databases listed in Table 2-2 were mostly extracted from the media. It can be also observed that the labeling of different types of emotion varies from database to database. This lack of unity makes any direct comparison between emotion recognition tests performed on different data bases very difficult.

The language is predominantly English, however there is a small number of corpora produced in German, Portuguese, Spanish, Chinese, Korean, Hebrew, Russian, Swedish, Dutch and Danish. The list presented is by no means comprehensive and currently new emotional speech corpora are being produced in a number of languages not listed in Table 2-2 [51].

Speech corpora used in this thesis contained naturally expressed stress and emotions; all speakers used English.

2.1.6 Language, culture and social environment based dependency of emotional expressions

Recent studies of emotional expression investigating cross-cultural [52] and socio-environmental differences between emotion portrayals aim to establish whether similar algorithms and features can be used to differentiate between emotional expressions across all languages and cultures. In [52] the first large-scale experiments were conducted collecting data from 9 languages on three different continents. The results provided moderate support for the claim that acoustic parameters of speech are in large parts controlled by universal psychobiological mechanisms. It was demonstrated that judges from different cultures, speaking different languages, recognize the expressed emotions with accuracy much higher than chance. The subjective emotion recognition was based on content-free utterances composed of phonological units from different Indo-European languages. It was suggested that the language-independent emotional cues are included in segmental information (fundamental frequency, formants) or suprasegmental structure
CHAPTER 2. REVIEW OF EXISTING METHODS OF STRESS AND EMOTION CLASSIFICATION IN SPEECH

(intonation, rhythm, timing). However, further studies confirming these findings are needed.

2.1.7 **Speaker based dependency of emotional expressions**

Although psychological studies indicate that certain speaker-differences in emotional expression exist, the majority of automatic emotion recognition studies are concerned with the speaker-independent approach [16, 51]. Relatively high correct recognition rates reported by these studies provide strong support for the claim that vocal expression of emotion is largely speaker-independent at least at the level of commonly used acoustic cues such as fundamental frequency, formants, energy, etc.

2.1.8 **Gender based dependency of emotional expressions**

There is a strong research evidence suggesting that the vocal emotional expressions conveyed by acoustic features differ between genders [51, 53, 54]. In [53] a gender specific selection of features was used to classify 5 different emotions (anger, happiness, neutral, sadness and surprise). The gender specific classification produced correct rate of 61.1% for male subjects and a corresponding rate of 57.1% for female subjects. Without the gender specific selection of features, a correct classification score of 50.6% was obtained. In [54], five emotional classes were considered (neutral, angry, anxious, dysphoric, and happy). Four different combinations of features derived from the Teager energy operator (TEO) and two different classifiers: Probabilistic Neural Network (PNN) and Gaussian Mixture model (GMM) were tested and compared. In all cases, the classification rates for females (38%-62%) were higher than for males (32%-53%).

Although, the issues of language, culture and person based differences of emotional expression are extremely valid and important from the perspective of emotional speech research, their scope is very large and goes beyond the scope of this thesis. The experiments presented in this thesis are limited to speaker independent recognition of emotions and only gender based differences are considered.
2.2 How stress and emotion are expressed in speech?

This section explains the thesis research rationale and provides a brief review of physiological and acoustical basis of stress and emotion recognition in speech.

2.2.1 Research rationale

The validity of an automatic stress and emotion detection in speech is based on the assumption that the stress or emotional state of a person affects the acoustic qualities of speech, and therefore the presence (or absence) of stress or emotion can be detected through an analysis of changes in the acoustical properties of speech. Moreover, a comparative analysis of acoustic characteristics of speech can provide information about the stress level or type of emotion expressed through speech.

2.2.2 Physiological basis

The physiological basis for the above research rationale result from a number of studies [55, 56] conducted since the 1980’s. These studies are showing that the effect of emotion on speech is due to tonic activation in the somatic nervous system and sympathetic as well as parasympathetic activation of the autonomic nervous system (ANS). For example, the temporal rate of articulation and the frequency range of vowels can be modified by changes in the timing of the muscle movements controlling the motion amplitudes of the articulatory structures. These muscles are driven by the limbic control system. Direct sympathetic and parasympathetic effects controlled by the limbic system can indirectly influence speech production by introducing changes in respiration and phonation. These could affect the fundamental frequency of vocal cords vibration (pitch) by causing changes in subglottal air pressure and vocal cord tension. Other ANS-caused effects such as mucus secretion may significantly alter speech characteristics during intense emotional arousal.

Psychomotor disturbances appear to be the unique and defining symptoms of mood disorders. There is evidence that symptoms of psychomotor agitation or retardation may
actually precede the full onset of a depressive episode and that these symptoms may be used to detect and monitor mental illness [30].

2.2.3 Acoustic properties of speech as indicators of stress or emotion

Emotional speech recognition aims to automatically identify the emotional or physical state of a person from his or her voice. The emotional state of a speaker produces what is called the emotional aspect of speech. Although, the emotional aspect of speech does not alter its linguistic contents, it is an important factor in human communication because it provides feedback information about physiological and mental state of a speaker. The concept of an automatic emotion or stress recognition was introduced in mid-1980s when a number of authors suggested the use of statistical properties of speech in automatic emotion recognition [57, 58].

2.3 Why is it important to study emotional speech?

This Section shows potential applications of emotional speech analysis and classification and explains possible benefits for the society and industry.

Just as effective human-to-human communication is virtually impossible without speakers being able to detect and understand each other's emotions, human-machine communication suffers from significant inefficiencies because machines cannot understand our emotions or generate emotional responses. Words are not enough to correctly understand the mood and intention of a speaker and thus the introduction of human social skills to human-machine communication is of paramount importance. This can be achieved by the researching and creating new models for speech that embrace the signal, linguistic and emotional aspects of communication. Emotional models of speech are needed to generate natural sounding synthetic speech, and provide the analytical means for improved machine understanding of human spoken language. The use of emotional speech technology will enable machines to gain human-like attributes such as:
anticipation, social plasticity, and intuition and make them capable of adapting their responses to the emotional states of users.

Speech technology is an important part of our daily lives. Mobile phones in Australia exceed the Australian population; increasingly a product support call is met by a synthetic voice and recognition system and synthetic, computer voices announce street names on satellite navigation devices. However, amongst all these applications, user complaints are still common and often surround the unnatural sound of a synthetic voice, the failure of a digitally coded telephone voice to sound 'normal' or the poor recognition performance of a telephone answer system as a user becomes increasingly frustrated. All of these problems can be traced back to a fundamental issue: currently, our models for human speech are severely limited by the fact that they do not describe the emotional content of speech signals.

2.3.1 Applications of emotional speech technology

The use of emotion in speech systems can be divided into:

− Transmission: to faithfully transmit a user's emotional speech through a telecommunications path;
− Synthesis: to create a natural sounding synthetic voice;
− Recognition: the need of a system to recognize users’ emotions from their speech;
− True emotional, linguistic software systems.

The degree to which emotional speech analysis and modeling are currently employed varies across the application areas of speech technology:

(1) Emotional speech synthesis

Three major approaches to emotional speech synthesis dominate the literature: formant synthesis, concatenative synthesis and unit selection [16]. Formant synthesis generates acoustic speech data entirely based on rules surrounding the acoustic correlates of the speech and does not utilize human speech recordings. Acoustic profiles for each emotion
category are derived from the literature and manually adapted [59] to create a signal; the resulting speech, has an unnatural, mechanical sound. In contrast, concatenative synthesis and unit selection [59] joins recordings of a human speaker to generate the synthetic speech. The result is a synthetic utterance with the same pitch and duration values as those of the actor.

(2) Emotion in compressed speech

Speech communication systems such as mobile telephone networks employ compression of the speech signal for efficient storage and transmission. Existing speech compression standards have been designed without considering the impact on the emotion conveyed to the listener. Recent results [60] have shown that speech compression can lead to inaccuracies in correctly communicating a person’s emotional state; this can be of critical importance in scenarios such as calls to public safety officers.

(3) Effect of emotion on speech and speaker recognition

The performance of existing speech and speaker recognition systems, which assume a noise-free environment, degrades rapidly in the presence of noise, distortion and speaker stress [61-63]. Literature in this area is very limited and speech modeling methods are confined to general pattern recognition techniques that use characteristic features (such as spectral parameters) and statistical modeling approaches.

(4) Intelligent virtual agents

A lot of works have been done to endow intelligent virtual agents with emotions [64-67], in order to facilitate their interactions with humans, not only by enhancing the capability of agent to correctly recognize speakers' emotional states, but also by improving the technologies of agent to emotionally express itself more efficiently through vocal intonation, expressive facial and body expressions. In Ginevra Castellano's work [68], the iCats robot, which can identify players' emotions and display facial expressions while playing chess with children, could dramatically increase the level of users' engagement and the interest.
(5) Emotion recognition for call centers

Another area of commercial applications of emotion and stress detection improving human-computer interaction is in a call-center environment [57, 58, 69, 70]. The affect recognition system could detect potential problematic points in the telephone conversation, caused by customers expressing anger and frustration in dealing with the automatic system. The call center could then interactively modify the dialog strategy or switch to a human operator for assistance. In general, the deployment of emotion detection system might improve customers’ satisfaction and help to deal with the customer disputes more efficiently. In addition, the detection of customers’ emotional states in telephone conversations could provide important feedback helping to improve services in the future.

(6) Emotion recognition in lie detection technology

In spite of the controversial validity [71-73], the affect detection systems still prove a potential application in a lie detection technology [74], which could recognize voice stress and emotion while the subject was in the response to a stimulus (questions). However, the interpretation of the intensity and type of expressed emotion (or stress) in relation to the question needs to be carefully examined and interpreted by qualified psychologists.

(7) Emotion recognition in psychology, medicine and behavioral science

The emotional research area is currently based around the two general groupings of psychological theories related to emotion expression.

The first group, represented by theoretical proposals made in [75], postulates the existence of distinctive patterns of prototypical emotions (e.g. anger, joy, sadness, fear). In this tradition, basic emotions are triggered by appropriate eliciting events that produce static emotion-specific response patterns.
The second group of theories includes componential appraisal models [76] which represent emotions as dynamic and adaptive processes which continuously evaluate events in terms of novelty, pleasantness and significance, and produce highly variable expressions comprised of emotional mixtures. Typically, it has been suggested [77, 78] that these emotional patterns and events are reflected in the fundamental frequency, formants, energy and timing parameters of speech.

The rapid development of digital signal processing and the deep exploration in the disciplines of psychology, linguistics, biology and medicine pave the wide application of affects recognition system in the field of behavioral and mental health sciences. Affect (emotion) based analysis of audiovisual recordings of patients engaged in conversations with doctors and/or family members became a standard research tool in behavioral psychology [29, 79, 80] and medical diagnosis of depression [30, 81-85], autism [86], Asperger syndrome [87], Alzheimer [88], schizophrenia and Parkinson's disease [89].

2.3.2 Benefits of the emotional speech research to the industry and society

Speech is the fundamental form of human communication, and much (if not all) human speech is the product of a speaker’s emotional state. However, our speech processing systems to date have lacked effective processing of that emotion. The development of emotional speech technology has a potential to provide significant benefits to the national and international industry and society in general.

(1) Benefits to the national security

In particular, the national security can benefit from forensic applications of emotion detection (new types of lie detectors, emotional speech analysis of suspects, terrorists, kidnappers, hostages). Public safety, border control and internet security can benefit from improved automatic speech and speaker recognition systems.
(2) **Benefits to the national safety**

The development of emotional speech technology will open possibilities of new applications such as automatic assessment of mental state of people working under high-risk and high-stress conditions that require an optimal mental and emotional state (e.g. heavy machinery operators, people working with dangerous chemicals, poisons and radioactive materials, construction workers, pilots, car and bus drivers surgeons) [91-98].

(3) **Benefits to the medicine and psychology**

Mental health and medicine will benefit from the development of automatic systems providing quantitative measures supporting diagnosis of emotional disorders such as: depression, Alzheimer, and autism is currently researching a diagnostic system for early detection of depression). Also, natural sounding synthetic speech capable of emotional expression will improve speech aids for mute people and new automatic training systems can be designed to teach autistic children how to recognize emotion in speech.

(4) **Benefits to the human-computer communication systems**

Education will benefit from more natural sounding synthetic speech that can be used in automatic tutorials applying text-to-speech technology. The Industrial and domestic sector will benefit from improvement of synthetic speech sounds used in automatic customer service systems, industrial and domestic robots, and human-computer communication.

Emotion analysis of customer speech will help to analyze efficiency and quality of automatic customer services.

(5) **Benefits to the entertainment and media industry**

Entertainment and media will benefit from improvements in the quality of computer games, and the possible automatic generation of voices in animated films and synthesis of actors’ speech in other languages.
2.4 What are the existing methods of an automatic stress and emotion recognition in speech?

This section introduces the basic principles of the pattern recognition methodology used in the automatic stress and emotion classification presented in this thesis.

In the early 1990s, progress in the computer technology made it possible to design new, advanced feature extraction and classification algorithms. Computers became capable of processing large number of data provided by hours of audio-visual recordings within practical computational times. These improvements opened ways to practical implementations of new important applications of the speech technology including speech and speaker recognition and more recently, stress and emotion recognition.

2.4.1 Levels of emotional speech analysis

Stress and emotion expressed in speech can be analyzed at three different levels [99, 100]:

- **Physiological level**: investigating the physiological parameters during the production of speech, describing nerve impulses or muscle innervation patterns of the major structures involved in the voice-production process.

- **Phonatory-articulatory level**: investigating physical properties of speech sounds and describing the position or movement of the major structures related to speech production mechanism.

- **Acoustic level**: dealing with acoustic aspect of speech signals, describing characteristics of the speech wave form emanating from the mouth, such as fundamental frequency, formants, duration of speech and so on.

The majority of the recent studies at the physiological and phonatory-articulatory levels require intrusive measurement as well as a high level of expertise.
This thesis is concerned only with the acoustic level of speech analysis, which does not require intrusive data collection methodology and relays on relatively low-cost and non-invasive acquisition of speech recordings.

### 2.4.2 A general framework of the automatic emotion and/or stress recognition in speech

Figure 2-2 illustrates a general flowchart of automatic speech stress and emotion recognition in speech. The diagram is based on the well established and widely used pattern recognition approach [101]. In the past, it was successfully applied in numerous speech [102] and speaker recognition [103] systems.

**Training**

<table>
<thead>
<tr>
<th>Speech with Known Class Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-processing</td>
</tr>
<tr>
<td>Voiced Speech Detection</td>
</tr>
<tr>
<td>Feature Extraction</td>
</tr>
<tr>
<td>Feature Selection</td>
</tr>
<tr>
<td>Modelling</td>
</tr>
<tr>
<td>Class Models</td>
</tr>
</tbody>
</table>

**Testing (Classification)**

<table>
<thead>
<tr>
<th>Speech from Unknown Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-processing</td>
</tr>
<tr>
<td>Voiced Speech Detection</td>
</tr>
<tr>
<td>Feature Extraction</td>
</tr>
<tr>
<td>Feature Selection</td>
</tr>
<tr>
<td>Classification</td>
</tr>
<tr>
<td>class label</td>
</tr>
</tbody>
</table>

**Figure 2-2** The general flowchart of the speech stress and emotion recognition and classification system including two stages: training and testing (classification).

The emotion/stress recognition method illustrated in Figure 2-2 consists of two stages: training and testing (or classification).

The training process is an iterative procedure which usually has a supervised character; the speech samples with known class labels (known emotions) are first pre-processed to reduce the noise and remove the silence or unvoiced intervals (see section 3.2.7). The labeled and pre-processed speech is then used to calculate sets of acoustic feature parameters characterizing each emotional class. In some cases the features may undergo a process of data reduction (or redundancy removal) through an optimal selection of features. During the modeling stage the characteristic features are used to derive class models in a form of probability density functions (pdfs), statistically describing each
class, or in a form of neural network structures with nodes represented by sets of constant weights derived from the training data.

Once the class models are generated, the classification process can be conducted. Speech samples from unknown classes are subjected to the pre-processing, feature extraction and feature selection procedures usually identical to those used during the training process. A classification method is then used to perform the pattern matching and decision making process and produce the most probable emotional label for the examined speech sample. The classification is usually based on a pattern matching approach that determines class that produces the highest pdfs’ values or activates certain nodes in a neural network structure.

The existing analytical models of speech production such as the source-filter model [104] and speech prosody models (e.g. the Fujisaki model [105]) do not take into account mechanisms explicitly responsible for generation of the emotional aspect of speech. It is usually assumed that stress and emotion alter somehow the classical parameters defined by these models (e.g. fundamental frequency, formants, energy and timing) and that the emotional cues can be determined through the analysis of these changes.

Due to the lack of general theories and models of emotional speech, the dominant approach to the emotion classification process (Figure 2-2), has a data-driven character. The class models are derived through a stochastic training based on a large collection of speech data rather than by calculation of knowledge-based parameters.

The main disadvantage of this approach is that its accuracy depends on the existence of a large collection of highly representative training data containing sufficient numbers of samples representing different emotional classes. Furthermore, the training process assures validity of testing only if the testing data is collected within the same or similar environment to that from which the training data was acquired. This leads to high sensitivity to any environmental changes (e.g. speech recording conditions, acoustic noise, differences in accent, differences in age of speakers, etc.).
2.4.3 Existing methods of feature extraction in emotional speech analysis

(1) Prosodic features: fundamental frequency $F_0$, formants, energy and rhythm

The concept of an automatic emotion recognition was introduced in mid-1980s when a number of authors suggested the use of statistical properties of speech in automatic emotion recognition [20]. The prosodic features: fundamental frequency, formants, energy and rhythm, were widely exploited for stress and emotion recognition, because they were related to the arousal level of emotions.

It was observed that, the prosodic features of speech produced under stress and emotion vary from the features under neutral condition. The most often occurring changed include changes in the utterance duration, decrease or increase of pitch, shift of formants and different levels of energy. For instance, Ververidis and Kotropoulos [51] have shown that anger is characterized by higher level of energy and pitch, compare to other four emotions: disgust, fear, joy and sadness. It was also shown that male speakers express anger with a slower speech rate than a female under the similar conditions [106].

Liscombe et al. [70] extracted 10 statistical analyzed prosodic features from energy, formants, pitch and the ratio of voiced frames to identify the non-negative (positive and neutral) and negative emotions. Similar feature sets were applied in [107], [108], [109], [110], [111-115] to identify different kinds of emotions.

Generally, the statistic prosodic features did not achieve high affection recognition accuracy. This could be partially attributed to the difficulty to calculate accurate values of the prosodic features, as well as the difficulty to estimate parameters based on the speaker and text independent basis. For example, a widely known autocorrelation method for pitch estimation was shown to be very sensitive to the presence of noise leading to the interference from the first formant. The formants are most often calculated using linear prediction (LP) analysis. Like in the case of the fundamental frequency, the most often occurring problem related to the LP formants estimation is the false identification due to the noise. The rate of speech could potentially provide information of different emotions,
but only assuming speaker dependency. Since the speech rate is strongly speaker-dependent, it generally does not provide consistent results in speaker-independent emotion classification. For instance, some speakers prefer express anger with an increasing speaking rhythm, though the others reduced the rate. These shortcomings as well as the fact that there are no universal consensus due to the speaker or textual dependency, limit the use of prosodic features (pitch, formants and rhythm) as efficient stress and emotion discriminates in speech. Due to the key importance of the feature extraction to the stress and emotion recognition, recent research efforts are predominantly focused on finding the right type of features and a large number of new approaches to the feature extraction have been proposed.

(2) Mel Frequency Cepstral Coefficients (MFCC)

The Mel Frequency Cepstral Coefficients (MFCC) used widely in speech and speaker recognition provide signal characteristics based on the human auditory perception characteristics and usually lead to improved performance compare to features that don’t take into account these characteristics. It has been therefore argued that feature extraction based on human auditory perception could also provide superior results in the cases of stress and emotion recognition.

The MFCCs have been applied to the stress and emotion classification in speech [112-117] leading to relatively moderate results.

Nwe et al. [118] applied the MFCCs to differentiate between six types of emotions (anger, disgust, fear, joy, sadness and surprise). An emotion database was specially designed and set up for text-independent emotion classification. A total of 12 speakers were employed to generate 720 utterances. The classification accuracy for six types of emotions using MFCC feature was 59%, which was better than the linear coefficients linear prediction cepstral coefficients (LPCC) which achieved 56.1% but not as good as 78% of accuracy achieved by the proposed short time log frequency power coefficients.
Iliev *et al.* [119] applied MFCCs features of different orders to the glottal waveform and to speech signal to classify 4 emotions: happy, angry, sad, and neutral. The database of the emotional speech was collected in an anechoic chamber. The results indicated that the best performance was provided by MFCC of order 6 for both glottal and speech signals, with the classification accuracy around 60%.

Clavel *et al.* [111] applied feature vectors including pitch, jitter, shimmer, formants and MFCCs to detect fear-type emotions occurring during abnormal situations. The SAFE corpus was developed in this work, based on the fiction movies. The corpus contains recordings of both normal and abnormal situation in order to classify fear and neutral emotions. Despite the diversity of the data, the system obtained a promising result with the correct recognition rate around 70%.

Yildirim *et al.* [120] used MFCCs features to detect three emotions (frustrated, polite and neutral) in children during spontaneous dialog interactions with computer characters. It was observed that MFCC features achieved the higher average correct classification rates, ranging from 66.4% to 70.6% across genders and age groups, compared to other acoustic features: pitch frequency, root mean square energy, voicing and zero-crossing-rate.

Zhou *et al.* [91] applied MFCC to do the pairwise stress classification using simulated (neutral, angry, loud and Lombard) and actual (neutral and actual) domain of SUSAS (Speech Under Simulated and Actual Stress) database. The results indicated that the proposed TEO based features achieved better recognition accuracy than the traditional features MFCC and pitch.

In general, most of the existing studies showed that the MFCCs achieve moderate levels of correct classification rates for emotions [91].

(3) TEO based features

Another important type of features proposed for emotion recognition by Zhou *et al.* [91] provides sensitivity to the number of additional harmonics and cross-harmonics in
speech signal occurring due to additional sound sources generated by a non-linear air flow in the vocal tract. It is assumed that in the emotional state or for stressed speech, the fast nonlinear air flow generates vortices located near the false vocal folds. The feature extraction method described in Zhou et al. [91] uses the area under the normalized Teager energy (TEO) autocorrelation envelope. The correct classification rates for two classes neutral and stress using SUSAS database were around 90% which was higher than rates achieved using the fundamental frequency F₀ and MFCC. Similar results were obtained in Gao et al. [121] TEO features were used to classify 4 emotions (happy, sad, angry and neutral) in Mandarin speech. The results demonstrated superior performance of the TEO based features when compared to the MFCC features.

Tin et al. [122] applied Teager energy operator (TEO) based nonlinear frequency domain LFPC features (NFD-LFPC) and TEO based nonlinear time domain LFPC features (NTD-LFPC). These features were tested using five stress categories in the simulated portion of SUSAS database. Results showed that the NFD-LFPC gave 86% of the correct classification rates, while the NTD-LFPC provided 76%.

In the work by Torabi et al. [123], a combination of TEO, pitch and LFPC features was tested in simulated stress classification (SUSAS database) providing on average about 89% of the correct classification rates.

In conclusion, the TEO based features appear to provide the highest correct classification rates in stress and emotion recognition. This thesis provides more detailed investigation into the TEO and related nonlinear features in Chapter 6.

(4) Glottal features

Several glottal features were studied as potential cues for emotion differentiation. It has been demonstrated that the glottal features show strong correlation with the emotional speech characteristics He et al. [124].

Moore et al. [81, 82] investigated the role of glottal features in the detection of the clinical depression, which is a mood or emotion disorder. The classification was based on
the following glottal features: glottal timing (closed and open instants of glottis), glottal ratios (ratio of closing phase to opening phase, ratio of open phase to the total cycle, ratio of closed phase to the total cycle, and so on) and glottal shimmer. The results indicated that the glottal features showed very good performance when used in combination with prosodic features (fundamental frequency and formants) providing 89% of correct classification rates for males and 92% for females.

Ozdas et al. [125] investigated the slope of the glottal flow spectrum achieving good classification rates (85%) for the classification into depressed and near-term suicidal patients.

Iliev et al. [126] applied glottal features representing glottal symmetry to classify happiness, anger and sadness, recorded from six subjects speaking ten sentences. The classification performance varied between 48.96% and 82.29% depending on the type of emotion.

Like TEO based features, the glottal features also appear to provide high classification rates. Experiments described in Chapter 6 of this thesis use glottal features in stress and emotion classification and point into the high importance of these features as efficient stress and emotion differentiators.

(5) Impact of different frequency bands on stress and emotion discrimination

Feature extraction based on limited frequency band could provide very useful data reduction increasing computational efficiency. Elimination of frequency bands that do not provide important emotion-characteristic information could also lead to improved classification rates.

The majority of investigators put high significance on the low frequency bands, such as the 0-1.5kHz [20, 30, 127], whereas others suggest the opposite [118].

The effect of different frequency bands on stress and emotion classification has been investigated in this thesis. More details can be found in Chapter 8.
2.4.4 Existing feature selection methods for emotional speech

In general, there are two distinct mechanisms for feature selection namely the filter methods and the wrapper methods [128]. Figures 2-3 and 2-4 show the filter and the wrapper approach respectively. The main difference between these two approaches is that the filter based feature selection does not take into account the classification results; the optimal selection of futures is based on a simultaneous increase of the inter-class separability measure and decrease of the intra-class separability measure. The wrapper method, on the other hand selects an optimal sub-set of features based on the maximization of the correct classification rates.

![Figure 2-3 Feature selection using the filter approach.](image1)

![Figure 2-4 Feature selection using the wrapper approach.](image2)

The filter methods are considered inferior to wrapper methods, however wrapper methods are computationally more demanding than filter methods. Both approaches suffer from the fact that the optimal selection applies to a particular set of training data and in general, there is no guarantee that the classification based on the test data not used during training phase will provide good results.

It has been widely acknowledged that irrelevant features may reduce the classification accuracy and the robustness of the classifiers. An optimal feature selection provides benefits for two reasons. Firstly, smaller and uncorrelated feature sets lead to generally
better performance [129]. Secondly, the exclusion of irrelevant or non-informative features reduces the classification time. Feature selection algorithms have been applied in the stress and emotion recognition systems showing performance improvements.

The principal component analysis (PCA) is a relatively simple and the most commonly applied filter type of feature selection. Grimm et al. [130] classified four emotions: happy, angry, neutral and sad and used the principal component analysis to reduce the dimension of acoustic features derived from pitch, speaking rate, intensity and spectrum.

In Ververidis and Kotropoulos [128], features related to the energy, pitch, and formants were extracted from the Danish Emotional Speech database [131]. The sequential forward selection method (SFS) was used to automatically select best performing features for each gender. The classification results based on were 50% for both genders, 61% for males and 57% for females. In the follow up study Ververidis and Kotropoulos [132] reported a small improvement of the classification rates when the sequential forward selection method (SFS) was replaced by the sequential floating forward selection (SFFS).

A sequential application of both the filter and the wrapper methods for an optimal feature selection can be found in Sedaaghi et al. [133]. An adaptive genetic algorithm search was used to select an optimal sub-set of features that maximized the correct classification rates. These features were then excluded from the sequential floating feature selection. The tests were performed on the Danish Emotional Speech database using features extracted from formants, pitch, and spectral energy and provided correct classification rates of 49% for both genders.

In Altun and Polat [134], four different filter based feature selection algorithms are compared in the task of emotion classification into 4 classes (anger, happiness, neutral and sadness) using the support vector machine (SVM) classifier. The best results were obtained using LSBOUND technique leading to about 80% of correct classification rates.
2.4.5 Existing classification techniques for emotional speech

The type of classifier used to generate class models and make the classification decision does not appear to have as high impact on the classification accuracy as the type of characteristic features.

Most studies of stress and emotion recognition use at least two classifiers to observe the consistency in the classification results and to choose the best classifier for a given type of characteristic features.

For example, Morrison et al. [69] applying utterances recorded from a real call center to classify anger, disgust, fear, happiness, sadness and surprise using seven different classification methods: SVM, MLP, KNN, K, RF, Stacking C and vote. The SVM, Stacking C and the vote methods showed the highest overall correct classification rates. All classifiers showed consistent classification accuracy ranging between 44.32% to 79.43%, depending on the type of emotion.

Pao et al. [135] used an emotion recognition system to compare several classifiers for clean and noisy speech. Five emotions, including anger, happiness, sadness, neutral and boredom were classified using Mandarin speech. The tested classifiers included: KNN, GMM and HMM. The experimental results showed that for clean speech KNN provided average accuracy of 72.2%, GMM-70.3% and HMM-62.5%. For noisy speech (SNR=5dB) the results were 26.5% for KNN, 38.7% for GMM and 35.5% for HMM, indicating that GMM classifier shows the highest robustness in the presence of noise.

The most often used classifier in the stress and emotion recognition task include: the k-nearest neighbor (k-NN) method, the support vector machine (SVM) classifier, the Gaussian mixture model (GMM), the hidden Markov model (HMM), and the artificial neural networks (ANN).

(1) k-nearest neighbor (k-NN) classifier
The k-nearest neighbor (k-NN) classifier is the simplest machine learning algorithm, which identify the object by the majority vote of its neighbors based on the distance (usually Euclidean). The training procedure of the k-NN classification algorithm requires only storage of feature vectors and class labels of training samples, which makes k-NN classifier one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data.

The k-NN method is a classification algorithm where the input feature vector is classified based on the class represented by the majority of the k-nearest feature vectors obtained during the training process. Given an input feature vector, the algorithm finds k closest feature vectors representing different classes. The class represented by the majority of the k nearest feature vectors is assigned to the input vector.

The disadvantage of KNN classifier is that its accuracy relies on the selection of an optimum number of neighbors and the most suitable distance measuring method.

In most studies, the k-NN classifier shows results consistent with other types of classifiers, and very close correct classification rates.

Shami et al. [136] compared the k-NN classifier with support vector machines (SVD) and Ada-boosted decision trees to identify different emotions using merged four databases: Kismet [137], BabyEars [138], Danish [24], and Berlin databases [26]. The results showed that all classifiers provided very similar results. It was also observed that the classifier performance depends on the type of features.

In a number of studies [112, 116, 119, 120], the KNN classifier also showed a good performance (comparable with GMM and HMM) in the stress and emotion classification.

(2) Support vector machine (SVM) classifier

In recent years, support vector machines (SVMs) have been widely used to solve binary classification problems. In a binary classification problem, a SVM constructs a hyperplane in a multidimensional vector space, which is then used to separate vectors that
belong to two different classes. A good separation is achieved by the hyperplane that has the largest distance to the nearest training vectors of each class. The two-class SVM method can be expended to a multi-class problem. It is usually done by reducing the single multi-class problem into multiple binary classification problems. Each of the problems yields a binary classifier, which is assumed to produce an output function that gives relatively large values for examples from the positive class and relatively small values for examples belonging to the negative class.

Although, the SVM classifiers give slightly better results than the GMM method, the computational cost is usually very high compare to GMM. A reduction of the computational time can be achieved by a parallel implementation of the SVM classifiers [139].

Examples of application of the SVM modeling to emotion classification can be found in [113, 116, 119, 140]. Zhang [141] compared a new fuzzy least squares support vector machines (FLSSVD) with the classical approach in speech classification into: anger, happiness, sadness and surprise. The new approach provided higher robustness in noisy conditions (lower SNR). For SNR-10dB the FLSVD provided correct classification rates of 60% whereas the SV gave only 35%.

(3) Gaussian mixture model (GMM)

The Gaussian mixture model (GMM) is a statistical technique modeling speech as a weighted sum of multivariate Gaussian probability density functions [103, 104, 142]. More detailed description of the GMM approach can be found in Section 3.2.2.

The GMM classifier has been widely used in the speech based pattern recognition system, and it showed a good performance in numerous studies of stress and emotion classification in speech [126, 143-145].

Zhou [146] applied the GMM in classification of Mandarin speech into five kinds of speech emotions: happiness, sadness, surprise, anger and neutral. The features included: MFCC, PLP, fundamental frequency and features based on Fisher's F-Ratio. Depending
on the type of emotion, the results provided about 70% to 95% of correct recognition rates.

In Tang et al. [147] a new boosted Gaussian mixture model (Boosted-GMM) method is compared with the traditional GMM approach in speech emotion recognition using four emotions: neutral, happy, sad and angry. The average correct recognition rates for the Boosted-GMM were around 90%, whereas the classical GMM provided about 85%.

Tao et al. [148] tested and compared different modelling methods for emotional speech including a linear modification model (LMM), a Gaussian mixture model (GMM), and a classification and regression tree (CART) model. The results showed that the LMM gave the worst results among the three methods. The GMM method was found to be more suitable for small training sets, while the CART method give the better results if trained with a large context-balanced corpus.

Iliev et al. [119] compared an optimum path classifier (OPF) with the SVM, GMM and k-NN classifiers in classification into four emotional classes: happy, angry, sad, and neutral. The features included the glottal symmetry parameters and MFCCs. Experimental results indicated that best performance was obtained for the glottal-only features with SVM and OPF generally providing the highest recognition rates, while for GMM or the combination of glottal and speech features performance was relatively inferior.

(4) Hidden Markov model (HMM)

The hidden Markov model (HMM) is a statistical modeling technique in which the modeling system is assumed to be a Markov process in which a current state depends on the previous state. The overall Markov chain represents the temporal structure (or pattern) of a feature parameter representing a given class.

Production of these patterns (or states) is represented by a probabilistic function \( \{b_{jk}\} \). The transition from a state i to state j is represented by probabilities [144]. The initial state of the system is given by the probabilistic vector \( \{p_i\} \). In other words, each class can
be described in a probabilistic way by defining values for elements of the three arrays: \( A = [144], B = \{b_{jk}\} \) and \( P = \{p_i\} \). In training phase, the probability arrays \( A, B \) and \( P \) are estimated for each class as a class model.

The class recognition phase in the HMM method involves computing the likelihood of generating the unknown input pattern using each of the stored model pattern and selecting that pattern which gives the greatest likelihood as the recognized pattern. This is known as the maximum likelihood classification.

The HMMs are well known in the field of temporal pattern recognition, and widely used in the speaker recognition, speech recognition and emotional speech classification system.

Nwe et al. [118] applied hidden Markov models to classify six categories of emotions - anger, disgust, fear, joy, sadness and surprise using short time log frequency power coefficients (LFPC) features in a speaker-dependent mode. Results showed an average accuracy of 80%.

Womack et al. [149] compared a single-channel HMM with a multi-channel mixture of HMMs to recognize four emotional states. The multi-channel approach yielded 62% of the correct classification rates, while the single channel HMM provided 10% lower rates.

(5) Artificial neural networks (NN)

Neural networks have been widely used in various pattern recognition problems; the strength of neural networks to discriminate between patterns of different classes has been exploited in a number of speech emotion recognition studies [114, 119, 150] showing moderate levels of success.

An artificial neural network (NN) is a nonlinear statistics data modeling tool. It is used to model complex relationship between the input and output data. A neural network consists of a group of interconnected artificial neurons, while their connection could be adapted through the change of the weight values during the learning process. There are
two major learning paradigms for neural network: the supervised learning which infer the mapping implied by the data pair \((x, y), x \in X, y \in Y\), while unsupervised learning was only given the data \(X\).

In [114] multi layer perceptron neural network (MLPNN) and a Random Forest classifiers were applied to recognize seven emotions (anger, happiness, anxiety/fear, sadness, boredom, disgust and neutral) using the Berlin emotional speech database [26]. In speaker dependent framework, artificial neural network classification reached an accuracy of 83\%, and Random Forest 77\%. In speaker independent framework, for artificial neural network classification a mean accuracy of 55\% was reached, while Random Forest reached a mean accuracy of 48\%.

In [150] a probabilistic neural network (PNN) was applied to classify eight emotions: anxiety, boredom, contempt, disgust, elation, angry, happiness, and interest. The average accuracy was about 58\%.

In Section 6.2.5 of this thesis experiments using the probabilistic neural network (PNN) in stress and emotion classification can be found.
3.1 Speech data used in this study

The speech data used for the purpose of this study was selected from two databases: the Speech under Simulated and Actual Stress (SUSAS) database and the Oregon Research Institute (ORI) database.

3.1.1 The Speech Under Simulated and Actual Stress (SUSAS) database

The Speech Under Simulated and Actual Stress (SUSAS) database [4] comprises a wide variety of simulated and actual stresses and emotions. The database is partitioned into five domains:

- Talking Style
- Single Tracking Task
- Dual Tracking Task
- Actual Speech under Stress and
- Psychiatric Analysis Domain

Only the first four domains are available in the current release of the SUSAS database. The speech was collected from 32 speakers (13 male and 19 female) within work environment (pilot cabin) and during a rollercoaster ride.

While performing these tasks, the speakers were reading a fixed list of 35 words presented in a random order. The list of these 35 words can be found in Table 3-1. The
speech corpora represented a narrow-band speech sampled by a 16 A/D converter with the sampling frequency of 8 kHz.

<table>
<thead>
<tr>
<th>Break</th>
<th>Change</th>
<th>Degree</th>
<th>Hot</th>
<th>East</th>
<th>Eight</th>
<th>Eighty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enter</td>
<td>Fifty</td>
<td>Fix</td>
<td>Freeze</td>
<td>Gain</td>
<td>Go</td>
<td>Hello</td>
</tr>
<tr>
<td>Help</td>
<td>Histogram</td>
<td>Destination</td>
<td>Mark</td>
<td>Nav</td>
<td>No</td>
<td>Oh</td>
</tr>
<tr>
<td>On</td>
<td>Out</td>
<td>Point</td>
<td>Six</td>
<td>South</td>
<td>Stand</td>
<td>Steer</td>
</tr>
<tr>
<td>Strafe</td>
<td>Ten</td>
<td>Thirty</td>
<td>Three</td>
<td>White</td>
<td>Wide</td>
<td>Zero</td>
</tr>
</tbody>
</table>

**Table 3-1 A vocabulary of 35 words used in the SUSAS database.**

**Speech under actual stress (SUAS) domain**

The speech corpora under actual stress from the Dual Tracking Task domain of SUSAS database [4] consists of speech recordings from pilots while performing a dual tracking task. The recording sessions protocol was similar to the protocol developed by Folds [151, 152] for the USAF School of Aerospace Medicine. It included a three-stage process designed such that during each stage the pilots were performing tasks under different level of stress. During the entire process, the pilots were aiming to achieve two key goals: flight control and target acquisition.

In the first stage, a pursuit task was released as a primary tracking task. Two parallel sinusoids were displayed on the computer screen, giving the appearance of a winding road which was scrolled down the screen over time. A stream of small circles was displayed on the screen as the response marker (output signal). The vertical positions of these circles were fixed at the center of the display, while the horizontal positions were determined by the movement of the control stick. As the roadway moved downward, the operators had to adjust the control stick to position the circles as close to the center of the road as possible.

The second stage was initialized immediately after the first stage; a secondary target acquisition task was added to the primary pursuit task. Two narrowly spaced vertical lines were drawn on the left portion of the screen, down along with small triangles between them. The vertical positions of the triangles were fixed at the center of the display, while a Gaussian distributed random value was added to the horizontal positions.
to move the triangles left or right. The pilots were required to move the second control stick to position the triangles back at the center of the two lines.

The final stage was initialized immediately after the second stage; the primary pursuit task was disabled and the pilots were performing only the secondary target acquisition task.

The task difficulty was controlled by time constraints for completion, and increased or decreased resource competition. To facilitate the collection of pilot’s speech, randomized words from the 35 word vocabulary listed in Table 3-1, were displayed on the screen and the pilot was instructed to read them quickly while performing the dual tracking task.

The speech recordings from the Speech under Actual Stress domain (SUAS) were collected from 7 speakers (3 females and 4 males).

In this thesis speech samples representing three different levels of stressful were used: high level stress, low level stressed and neutral speech.

- **High level stress:** For the high workload case, the pilots performed all the three task stages in 40 seconds.
- **Low level stress:** For the moderate workload case, the pilots finished all the three tasks in double time constraint of 80 seconds.
- **Neutral:** The neutral speech recordings were produced by speakers while seating in a quite, sound resistant chamber and not performing any tasks apart from reading the words.

The SUAS data used in this thesis contained a total of 3179 speech samples including 1202 recordings representing the high stress, 1276 recordings representing the moderate stress and 701 recordings representing the neutral speech.

In order to test the efficiency of stress classification based not only on random words but also on specific linguistic units such as vowels, the SUAS recordings were divided into the following sub-sets:

- words containing a single vowel ē,
words containing a single vowel ā,
- words containing different but single vowels, and
- all SUAS words (with single and multiple vowels).

The sizes of these sub-sets for different stress levels are listed in Table 3-2. This particular and rather limited selection of vowels was determined by a relatively small number of words and speech recordings available in the SUAS domain. However, on the positive side, the vowels ē and ā, are the two most commonly used vowels in English [239].

Table 3-2 Description of the Speech under Actual Stress datasets; H-high level stress, L-low level stress, N-neutral.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Vocabulary set</th>
<th>Number of Recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words with single ē vowel</td>
<td>break, change, eight, eighty, gain, strafe</td>
<td>133  143  59</td>
</tr>
<tr>
<td>Words with single ā vowel</td>
<td>break, change, east, eight, fix, freeze, gain, go, help, hot, mark, nany, no, oh, on, out, point, six, south, stand, steer, strafe, ten, three, white, wide</td>
<td>206  220  121</td>
</tr>
<tr>
<td>Words with mixed single vowels</td>
<td>break, change, east, eight, fix, freeze, gain, go, help, hot, mark, nav, no, oh, on, out, point, six, south, stand, steer, strafe, ten, three, white, wide</td>
<td>871  931  523</td>
</tr>
<tr>
<td>Words with single and multiple vowels (Speech Under Actual Stress (SUAS))</td>
<td>break, change, degree, hot, east, eight, eighty, enter, fifty, fix, freeze, gain, go, hello, help, histogram, destination, mark, nav, no, oh, on, out, point, six, south, stand, steer, strafe, ten, thirty, three, white, wide, zero</td>
<td>1202 1276 701</td>
</tr>
</tbody>
</table>

A sub-set with words containing a single ē vowel included 133 high-stress recordings, 143 low-stress recordings and 59 neutral speech recordings. A sub-set with words containing a single ā vowel included 206 high-stress recordings, 220 low-stress recordings and 121 neutral speech recordings. Finally the full set of SUAS words with single and multiple vowels included 871 high-stress recordings, 931 low-stress recordings and 523 neutral speech recordings.
3.1.2 The Oregon Research Institute (ORI) database

Depression is one of the most common mental health problems in the western world, and it is characterized by changes in emotional characteristics of patients. Most frequently, it is manifested by prolonged periods of sadness and withdrawal. In the case of the bi-polar depression, the prolonged periods of sadness are alternated by euphoric states of exaggerated happiness.

A number of studies in the area of child psychology indicate that the family relationships and the interaction processes between the family members are critical factors in the development and maintenance of depressive symptoms in adolescents. Sheeber et al. [29, 79] suggested that the quality of family interactions is relevant for understanding the development of depressive symptoms in adolescents. An adverse family environment is associated with depressive symptoms of adolescents. It has been reported that depression is inversely related to the level of support, attachment, and approval adolescents experience in the family environment.

In order to study the effects of family interactions on the adolescents’ behavior and emotional characteristics, the Oregon Research Institute recruited 716 parents and their offspring attending six different high schools in the Western Oregon to collect an audiovisual database. In total, 156 families participated fully in the year 1 assessment concluding completion of questionnaires as well as participation in the video-recorded problem-solving interactions by all family members. A slightly smaller number of 151 families out of the original 156 fully participated in the year 2 assessments consisting of both, the questionnaires and the problem-solving interactions.

The recordings were made using in a quiet laboratory room at ORI. Family members were seated a few feet apart. Lapel wireless microphones (model: Audio Technica ATW-831-w-a300) were placed on the participants shirts at the chest level in the way that did not impede speech behaviour.
The soundtrack of the video recordings from the problem solving interactions was used in this thesis to select speech samples for the emotion recognition experiments [29, 79, 80].

The ORI videotapes were annotated in real time by trained psychologist using the Living in Family Environments (LIFE) coding system [153]. The coded script provided a second by second, 5-digit coded description of: subject (one digit), content (two digits), recipient (one digit), and subject’s affect (one digit). The following class discrimination rules regarding the speech signal were used during the annotation process [153].

- **Angry**: this affect communicates displeasure. An angry person sounds like he/she is “fed-up”. Voice is lowered or raised beyond the limits of normal tone; words usage is abrupt, with one word or syllable being more strongly stressed; short clipped speech; irritation, annoyance, frustration evidenced by changes in the rhythm of speech and the way certain words are stressed.

- **Anxious**: this affect communicates anxiety, nervousness, worry, fear or embarrassment. The voice can be described as tense, fearful, concerned, startled, shocked, hysterical, afraid, uneasy, worried. There could be speech difficulty, stuttering, or slips of the tongue. Mouth and lips may tremble.

- **Dysphoric**: this affect communicates sadness and depression. Persons that are depressed may appear detached from the ongoing activity, tend to speak slowly, and use a low voice tone. The voice is sad, glum, distressed, discontented, withdrawn, discouraged, despondent, joyless, gloomy or melancholic. There is a low voice tone and slow pace of speech, often with sighing and yawning.

- **Neutral**: this affect occurs when the participant is relatively even-tempered, composed, instructional or reasonable. Neutral is described as a dividing line between negative and positive affects and is generally non-emotional. A neutral voice tone is even, relaxed, without marked stress on individual syllables. In situations where the person’s behavior contained a mixture of neutral and any
other affect category, the other category was coded. The voice quality is even-
tempered, without a trace of dejection, sternness, or sulkiness.

- *Contempt*: this affect communicates a lack of respect for the recipient. It suggests
that the speaker feels somewhat superior to the recipient and mocks or mimics the
recipient in a cold or rude way. The voice tone is sarcastic, insulting, cold,
condescending and disgusted.

- *Pleasant*: this affect reflects an interested and engaged quality. It is the bridge
between neutral and happy, or neutral and caring. It also reflects a change in
energy from passive to active listening. Subject becomes focused on the speaker
and genuinely engaged in the conversation. The voice tone is agreeable, receptive,
attentive, engaged, interested, amiable and cordial.

- *Happy*: this affect reflects mood that can be described as glad, silly, playful,
funny, hopeful, thrilled, pleased, excited exuberant, cheerful, delighted or
enthusiastic. The speech is full of laughter, giggling or smiling. The voice tone is
a high pitched or sing-song but not whining. Speech is faster or louder than usual
but not angry.

With the help of the Adobe Pro software, the soundtrack was extracted from the
original video files with the sampling frequency of 44100Hz.

The stress and emotion experiments presented in this thesis include experiments
conducted on the narrow bandwidth speech produced by down-sampling the original
recordings to the sampling frequency of 8kHz (4kHz bandwidth) as well as experiments
with a wide bandwidth produced by down-sampling the original recordings to the
sampling frequency of 22kHz (11kHz bandwidth). The narrow-band experiments are
described in Chapters 4, 5 and 6, and the wide-band experiments can be found in Chapter
7.

Testing of the narrow band was motivated by a strong interest in our technology
coming from the telecommunication industry which currently uses a narrow-band speech.
Using a narrow-band speech also provides an additional advantage of a reduced data size making real-time implementations of presented here methodology more realistic. In addition, having both data (SUAS and ORI) sampled at the same frequency helped to draw direct comparison conclusions between stress and emotion recognition results.

However, in order to observe the importance of different frequency bands on the emotion classification results, a wide-band speech was also tested in Chapter 7.

Audio files of 19 fathers and 20 mothers of the depressed adolescents, and 25 fathers and 7 mothers of the non-depressed adolescents were selected for the purpose of this study.

For each of the seven types of emotions (contempt, angry, anxious, dysphoric, neutral pleasant, and happy), 200 utterances were selected from the audio files for this research. These utterances included 50 utterances representing mothers of depressed adolescents, 50 utterances representing fathers of depressed adolescents, 50 utterances representing mothers of non-depressed adolescents, and 50 utterances representing fathers of non-depressed adolescents. The average duration of each utterance was 1.5 seconds.

Experiments described in Chapters 4, 5 and 6 provide classification into 5 emotional classes: angry, anxious, dysphoric, neutral and happy using a narrow-band speech, whereas, experiments described in Chapter 7 provide classification into 7 emotional classes: contempt, angry, anxious, dysphoric, neutral, pleasant and happy using a wide-band speech.

3.2 Stress and emotion classification system used in this study

Figure 3-1 illustrates a general flowchart of the automatic stress and/or emotion recognition system used in this study. The system consisted of a two-stage processing.

In the first stage called the training, the characteristic features representing known emotions (or stress levels) were used to train the emotional class models.
In the second stage called the classification (or testing), the characteristic features from speech samples of unknown classes were compared with the class models, developed during the training stage, to determine stressful and emotional classes to which they belonged.

The feature parameters examined in our experiments included the author’s new feature parameters as well as a set of classical features used to provide a reference point for a comparative assessment of the new features.

The classical feature extraction methods included:
- fundamental frequency $F_0$,  
- first three formants (F1, F2 and F3), 
- Mel Frequency Cepstral Coefficients (MFCCs) and 
- glottal time/frequency domain features.

The author’s new features included:
- parameters extracted from speech sub-bands of spectrograms without anisotropic filtering (SS-CB-AE, SS-BARK-AE, SS-ERB-AE),  
- parameters extracted from sub-bands of speech spectrograms with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE, SS-AF-ERB-AE),  
- parameters extracted from speech spectrograms and combined with a single log-Gabor filter and an optimal feature selection (SS-LGF-OFS),  
- parameters extracted from speech spectrograms and combined with a log-Gabor filter-bank, averaging and an optimal feature selection (SS-ALGF-OFS),  
- parameters extracted from speech spectrograms and sigma-pi units (SS-sigma-pi),  
- parameters extracted from the wavelet packet arrays combined with a log-Gabor filter-bank, averaging and an optimal feature selection (WP-ALGF-OFS),
Feature generation using spectrogram patches and log-Gabor filters (SS-SP-ALGF-OFS),
parameters derived from the Empirical Mode Decomposition (EMD) combined with calculation of an averaged Renyi’s entropy (EMD-AER),
features derived from the Teager Energy Operator (TEO) applied to the speech signal, (TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S),
features derived from the Teager Energy Operator (TEO) applied to the glottal waveform (TEO-PWP-G),
features derived from the spectral energy of speech (AUSEES)
features derived from the spectral energy of the glottal wave (AUSEEG)

Two well established methods of data modeling and classification were tested:
Gaussian Mixture Model (GMM) and
K-Nearest Neighbors (KNN).

The classification tests had speaker, text and gender independent character. The speech samples tested in Chapter 4, 5 and 6 were sampled at the 8 kHz rate resulting in the narrow-band of 4 kHz, and in Chapter 7 were sampled at the 22 kHz rate resulting in the wide-band of 11 kHz.

3.2.1 Pre-processing

(A) De-noising

Both SUSAS and ORI database were recorded in real-life noisy conditions. To reduce the background noise, a wavelet-based method developed by Donoho [154] was applied. Speech signal of length N and standard deviation \( \sigma \), was decomposed using the wavelet transform with the mother wavelet db2 up to the second level, and the universal threshold was applied to each wavelet sub-band. The signal was then reconstructed using the inverse wavelet transform (IWT).
(B) Silence removal for the ORI database

The SUAS data contained recordings of single words; therefore there was no need to remove silence intervals occurring between words or sentences. The ORI database, on the other hand contained sentences and therefore the silence intervals had to be removed prior to further processing.

The silence removal was applied to the ORI database on the frame-by-frame basis. Frames of length 256 samples with no overlap were used. The sampling signal energy was normalized for each frame using:

$$\tilde{E}[K] = (E[k] - \bar{E}[k])/\text{std}(E)$$

(3-1)

Where k is the frame number, $$\bar{E}[k]$$ is the mean value of energy for the given frames and std(E) is the corresponding standard deviation. The silence samples with energy less than zero were removed and the remaining speech samples were concatenated.

All de-noised and silence removal speech samples were inspected by listening to ensure that the de-noising process preserved the stressful and emotional aspect of speech.

3.2.2 Classification techniques

All of the feature extraction methods introduced in this thesis were tested in combination with the Gaussian Mixture Model (GMM) and the k-Nearest Neighbors (KNN) classifiers. In Chapter 6 selected features were also tested in combination with two other classifiers namely the Multileayer Perceptron neural Network (MLPNN) and the Probabilistic Neural Network (PNN). The following Sections provide brief descriptions of the classification techniques used in this Chapter.

(A) Gaussian Mixture Model (GMM)

The Gaussian Mixture Model (GMMs) [103, 104, 142] is a Bayesian feature modeling and classification algorithms widely used in the speech-based pattern recognition, since it can smoothly approximate a wide variety of density distributions.
The probability density function (pdf) drawn from a GMM is a weighted sum of $M$ component densities given as:

$$p(x | \lambda) = \sum_{i=1}^{M} p_i \cdot b_i(x)$$  \hspace{1cm} (3-2)$$

Where $M$ is the number of mixtures for GMM (in this thesis, $M$ is experimentally set as 5, which gives the best classification accuracy). $x$ is a $D$-dimensional random vector, $b_i(x)$ $i=1…M$ are the component densities and $p_i$ $i=1…M$, are the mixture weights.

Each component density is a $D$-variate Gaussian function of the following form:

$$b_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma|^i} \exp\left(-\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right)$$  \hspace{1cm} (3-3)$$

Where $\mu_i$ is the mean vector and $\Sigma$ is the covariance matrix. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} p_i = 1$. Each class is represented by a mixture model and is referred by the class model $\lambda_i$.

The complete Gaussian mixture model $\lambda_i$ for a class $i$, is the collection of the mean vectors, covariance matrices and mixture weights from all components densities. It can be written as the following set:

$$\lambda_i = \{p_i, \mu_i, \Sigma_i\}, i = 1,\ldots,M$$  \hspace{1cm} (3-4)$$

The Expectation Maximization (EM) algorithm was used to iteratively derive class models. The EM algorithm initialized with the class models $\lambda_i$ was estimated at each iteration for the new models $\hat{\lambda_i}$ while $p(x | \hat{\lambda_i}) \geq p(x | \lambda_i)$.

Then, for each test utterance $x_{test}^k, k = 1,2…M$ , (where $M$ is the number of test utterances), the $x_{test}^k$ was calculated and the probability density $p(x_{test}^k | \lambda_j), j = 1,2…N$ of each class model given the input features was estimated using the GMM. The classification task was then to choose the class with the highest probability $p(x_{test}^k | \lambda_j)$, given as:
\[ P(x_k^{\text{test}} | \lambda_j) = \frac{P(x_k^{\text{test}} | \lambda_j) P(\lambda_j)}{P(x_k^{\text{test}})} \quad (3-5) \]

where \( P(\lambda_j) \) was the a priori probability of class \( \lambda_j \) being the source of the test features \( x_k^{\text{test}} \) and \( P(x_k^{\text{test}}) \) was the probability of the test features being selected from the whole set of test features.

Since both probabilities \( P(\lambda_j) \) and \( P(x_k^{\text{test}}) \) were constant and known, the classification problem was to find the class model \( \lambda_j \) which maximized the pdfs \( p(x_k^{\text{test}} | \lambda_j) \) calculated by the GMM method. The test utterance was then assigned to the class with the highest pdf value.

**B) K-Nearest Neighbours (KNN) classifier**

The k-Nearest Neighbors (KNN) classifier [155] is one of the most fundamental classification algorithms, and it often is the first choices in classification problems when there is little or no prior knowledge about the distribution of the data.

The k-Nearest Neighbors algorithm is a supervised classification method in that it uses the known class labels of the training data. During the testing process, given an input (testing) feature vector, the squared Euclidean distances between the input vector and the training vectors are calculated as:

\[ d(TE_i, TR_l) = \sqrt{(TE_{i1} - TR_{l1})^2 + (TE_{i2} - TR_{l2})^2 + \ldots + (TE_{ip} - TR_{lp})^2} \quad (3-6) \]

where \( p \) represents the dimension of the feature vectors, \( TE_i \) is the tested feature vector, \( TR_l \) (\( l=1,2,\ldots,n \)) are the training vectors and \( n \) is the total number of the training vectors.

For each test sample \( TE_i \), the \( k^{\text{th}} \) minimum distances between \( TE_i \) and the training samples are determined from the distance set \( D \) given as:

\[ D = \{ d(TE_i, TR_1), d(TE_i, TR_2), \ldots, d(TE_i, TR_n) \} \quad (3-7) \]
The class represented by the majority of the k nearest samples is assigned to the test sample [156].

(C) Multilayer Perceptron Neural network (MPNN)

Artificial neural networks are widely used in the field of pattern recognition [101, 157, 158]. A number of studies reported their use in the emotion recognition in speech [159-163]. The main advantage of neural networks is that they can map complex and nonlinear relationships between the inputs and the outputs.

One of the most versatile and often used neural networks is the multilayer perceptron neural network (MLPNN) [164, 165]. Typically the MLPNN consists of the input layer, one or more hidden layers, and the output layer (see Figure 3-2).

![Multilayer Perceptron Neural Network](image)

**Figure 3-2** An example of a multilayer perceptron neural network (MLPNN) structure used in the training and testing stages for five emotions classification (happy, anxious, neutral, angry and dysphoric) in speech.

The MLPNN is a feed-forward network mapping a set of input vectors onto a set of outputs. The outputs are usually given in a binary form as illustrated in an example in Figure 3-2. Each hidden node is assigned a weight value $w_i$ and the output nodes produce...
binary outputs calculated as a sigmoidal function of the weighted sum of the outputs from the hidden layer.

The MLPNNs are usually trained by a learning algorithm. It calculates the errors between classes indicated by the output nodes and the classes represented by the training data and uses it to adjust the weights in the way that minimizes the errors. The adjustment of weights is usually guided by an optimization algorithm [166] as illustrated in Figure 3-3.

![Figure 3-3](image_url) An iterative training of the multilayer perceptron neural network (MLPNN).

In the experiments described in Chapter 6, MLPNNs with one hidden layer and the number of output nodes depending on the number of stress or emotion classes were used. The number of hidden layers was empirically calculated as $n/10$, where $n$ was the number of training vectors. A standard supervised training procedure was applied to derive an optimal set of hidden layer weights by an iterative minimization the network response error. The optimal set of weights was then used in the classification process.

A general flowchart of the training procedure is illustrated in (see Figure 3-2). The network error was calculated based on the responses of the output nodes. Each node was expected to respond with 1 if the input feature vector belonged to the class represented by this node and with 0, if the input vector belonged to any other class.
The experimental results described in Chapter 6 were obtained using three different optimization procedures: the Fletcher-Powell conjugate gradient, the scaled conjugate gradient and the resilient back-propagation. All of these algorithms represented a greedy, gradient-descent type of optimization and required calculation of the objective function derivative. The derivative-based algorithms are known to converge faster than the non-derivative procedures, however, they can easily get stuck in local minima providing sub-optimal solutions. For this reason their performance often depends on a “good” starting point (or initial guess) [166].

(D) Probabilistic Neural Network (PNN)

The probabilistic neural network (PNN) was introduced by Donald Specht [167]. Due to its efficiency, it quickly became one of the most prominent and widely used types of neural networks.

The main advantage of the PNN is that the training time is much less than that of the MLPNN, it is simple to implement, and it has an inherently parallel character.

The probabilistic neural network is a feed-forward algorithm which is comprised of three layers:

- the input layer,
- the radial basis layer which calculates the distances between the input vector $\mathbf{x}$ and all the training vectors $\mathbf{x}_{ij}$ and
- the competition layer which calculates the pdfs $f_j (\mathbf{x})$ for all classes and finds the class with the maximum pdf value.

Unlike the MLPNN, the PNN is based on statistical principles derived from Bayes’ decision strategy and non-parametric kernel based estimators of probability density functions. It has been shown [167] that the PNN is guaranteed to approach the Bayes’ optimal decision surface provided that the class probability density functions are smooth and continuous. The PNN uses the Parzen probability distribution function (pdf) estimators [101, 164, 165].
During the classification process, for a given test vector \( \mathbf{x} \), and for each class \( i \), the PNN calculates a sum \( f_i(\mathbf{x}) \) of the multivariate spherical Gaussian radial basis functions centred at each training vector using the following formula:

\[
f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sigma^pM_i} \sum_{j=1}^{M_i} \exp\left(-\frac{(\mathbf{x} - \mathbf{x}_{ij})^T(\mathbf{x} - \mathbf{x}_{ij})}{2\sigma^2}\right)
\]

(3-8)

Where \( i \) is the class number, \( j \) is the pattern number, \( \mathbf{x}_{ij} \) is the \( j \)-th training vector from class \( i \), \( M_i \) is the number of training vectors in class \( i \), \( p \) is the dimension of vector \( \mathbf{x} \), \( \sigma \) is the standard deviation (or smoothing factor). The value of \( f_i(x) \) represents the probability density function estimate for the \( i \)-th class. The classification decision is made in accordance with the Bayes’ rule stating that the test vector \( \mathbf{x} \) belongs to the class \( i \), if \( \forall f_i(\mathbf{x}) > f_k(\mathbf{x}) \) [168].

### 3.2.3 Average Percentage of Identification Accuracy (APIA%)

The cross-validation method was used to test the classification results [169]. According to this method, the total set of speech samples was randomly divided into the training and testing sets. This procedure was repeated several times, and the classification score was calculated as an average percentage of identification accuracy (APIA%) defined as follows [103]:

\[
APIA\% = \left(\frac{1}{N_r} \sum_{i=1}^{K} \frac{N_{Ci}}{N_{Ti}}\right) \times 100\%
\]

(3-9)

Where \( N_{Ci} \) is the number of test inputs correctly identified during the \( i \)-th trial, \( N_{Ti} \) is the total number of test inputs, and \( N_r \) is the number of trials (number of times the training and testing processes were repeated).
Chapter 4. Stress and Emotion Classification Using Classical Features

This Chapter describes stress and emotion classification experiments based on classical features including: fundamental frequency $F_0$, formants ($F_1$, $F_2$ and $F_3$), mel frequency cepstral coefficients (MFCC) and glottal features. The classical features represent parameters describing a classical source-filter model of speech production.

4.1 Classical source-filter model of speech production

Speech is the most important means of communication among humans. Apart from conveying linguistic information between speakers, it also carries a large paralinguistic content including vital information about speakers’ emotions, personalities, attitudes, feelings, levels of stress and current mental states. As a biological signal, speech contains a lot of medical diagnostic information and psychological behavioral information, which in comparison with other biological signals, such as for example ecg or eeg, has been very much under-utilized.

One of the reasons for this under-utilization of the vital information present in speech is the combination of high complexity and a wide bandwidth of the speech signal, which makes the analysis relatively more complex than in the case of other bio-signals. Another limiting factor is a serious lack of proper modeling and understanding of the speech production process.

The classical source-filter model has a linear character and was generated a few decades ago for the purposes of telecommunication engineering, where conveying of an accurate linguistic content was of primarily importance. It does not include mechanisms explicitly responsible for the generation of the paralinguistic aspect of speech. As a result the majority of the current approaches to emotional speech analysis rely on the assumption that the emotional state of a speaker affects in some way speech parameters assumed by the existing source-filter model. Subsequently these parameters including the
fundamental frequency $F_0$, formants and energy, or parameters derived from them, are the most often cited in the literature as characteristic features used in emotion recognition from speech [51].

The classical source-filter theory of voice production assumes that the air flow through the vocal folds (source) and the vocal tract (filter) is unidirectional and has a laminar character (Figure 4-1) [170, 238]. During phonation, the vocal folds vibrate. One vibration cycle includes the opening and closing phases in which the vocal folds are moving apart or together, respectively. The number of cycles per second determines the frequency of the vibration, which is subjectively perceived as pitch or objectively measured as the fundamental frequency $F_0$. The sound is then modulated by the vocal tract configuration and the resonant frequencies of the vocal tract are known as formants.

![Classical model of speech production assuming laminar flow of air through the vocal folds.](image)

**Figure 4-1** Classical model of speech production assuming laminar flow of air through the vocal folds.

In general, the human speech production system (Figure 4-2) consists of three elements: a power source, a sound source and sound modifiers. These three elements provide the basis of the source-filter model of speech production.
The power source results from the compressive action of the lung muscles, forcing the air to flow from the lungs towards the vocal tract and mouth.

The sound source: there are two types of speech sound: the voiced speech and the unvoiced speech.

- The voiced speech is produced by the vibrations of the vocal folds (voiced excitation). The vocal cords are tensed together and they vibrate as the air pressure gradually builds up forcing the glottis to open and then subsides when the air passes throughout the glottis. This process is periodically repeated producing a train of approximately triangular pulses. It has the frequency spectrum that contains the fundamental frequency $F_0$ and higher harmonics. The amplitude of the spectrum decays at a rate of approximately -12 dB/octave. The range of $F_0$ values for an adult male is 50Hz to about 250Hz with an average value of 120Hz. For an adult female the average is about 225Hz and the upper limit at 500Hz. For children, the average $F_0$ is 265Hz.

- The unvoiced speech is produced by a turbulent air flow through the constrictions of the vocal tract (unvoiced excitation). In the production of the unvoiced speech, the vocal folds do not vibrate. This turbulent air flow produces acoustic noise that is essentially white and can therefore be considered as having a flat continuous spectrum.

The sound modifiers for the voiced and unvoiced speech are: the vocal tract cavities (the oral cavity and the nasal cavity), lips, jaw, tongue and velum. Movement of those elements changes the shape of the vocal tract which leads to changes of its acoustic properties namely the resonant frequencies called formants. The vocal tract works as a time-varying pass-band filter with an infinite number of resonant frequencies and amplitudes constant for each of those bands. The process of lip radiations works effectively as a high-pass filter with a continuous spectrum and amplitudes increasing with the rate of +6dB/octave across frequencies [170, 238].
The overall average spectrum of the voiced speech shows an amplitude decay of amplitude with frequency at an average ratio of -6dB/octave. The overall average spectrum of the unvoiced speech shows an increase of amplitude with frequency at an average ratio of +6dB/octave.

The following Sections describe experiments of stress and emotion classification in speech using classical features such as: fundamental frequency $F_0$, formants, mel frequency cepstral coefficients (MFCC) and glottal features.

### 4.2 Calculating fundamental frequency ($F_0$) and formants

As illustrated in Figure 4-3, for each frame of the voiced speech, the fundamental frequency $F_0$ of the vocal folds vibration was estimated simultaneously in the time domain using the autocorrelation method [170], and in the frequency domain using the cepstral method [170]. The average value of these two measurements provided the final estimate of $F_0$.

The frequency domain cepstrum method of the $F_0$ estimation looked for a periodicity in the log spectrum of the signal; if the log amplitude spectrum contained many regularly...
spaced harmonics, then the Fourier analysis of the spectrum was expected to show a peak
 corresponding to the spacing between the harmonics: i.e. the fundamental frequency.

![Diagram of fundamental frequency estimation](image)

**Figure 4-3** A flowchart of the fundamental frequency estimation method.

The time domain autocorrelation method, on the other hand, estimated the fundamental
 frequency directly from the waveform using the autocorrelation function which was
 expect to show peaks at delays corresponding to multiples of the glottal wave periods
 \( 1/F_0 \).

The first three formant frequencies \( F_1 \), \( F_2 \) and \( F_3 \), where estimated as the resonant
 frequencies of the vocal tract filter using the linear predictive (LP) analysis [83]. Based
 on the source-filter theory, the vocal tract was modeled as a LP recursive all pole IIR
 filter. The formants values were obtained by factoring the predictor polynomial, and
 solving the roots of the polynomial to find the locations of the resonances representing
 the values of formants.

### 4.3 Calculating the Mel-Frequency Cepstral Coefficients (MFCCs)

The mel-frequency cepstral coefficients [171, 172] are widely used as acoustic features
 for speech pattern recognition.

The mel-frequency cepstral coefficients are calculated for the outputs from the bank of
 auditory filters equally spaced on the logarithmic frequency scale called the mel scale.
 The mel-scale frequencies can be calculated as follows:
Where \( f_{\text{mel}} \) represents the frequency on the logarithmic mel scale corresponding to the frequency \( f_{\text{Hz}} \) given in Hz on the linear scale.

Since, the mel scale provides closer approximation of the human auditory system’s responses \(^{173}\) than the linear frequency scale, it is frequently expected that feature parameters calculated across the logarithmic frequency scale should provide better defined as features based on the linear scale.

The MFCCs were derived using the following steps:

**Step 1:** Assuming that \( y[n] \) (\( n=1,\ldots,M \)) represents a frame of speech which was pre-emphasized and multiplied by a Hamming window, \( y[n] \) (\( n=1,\ldots,M \)) were transformed into the frequency domain using \( M \)-points discrete Fourier transform (DFT), and calculate the energy spectrum samples \( |Y[k]|^2 \) (\( k=1,\ldots,M \)) as:

\[
|Y[k]|^2 = \left| \sum_{n=1}^{M} y[n] \cdot e^{-\frac{j \cdot 2 \pi k n}{M}} \right|^2
\]

**Step 2:** The spectral energies \( |Y[k]|^2 \) were mapped onto the logarithmic mel scale using a bank of triangular overlapping filters. The filter response \( \psi_i[k] \) of the \( i \)-th filter in the bank was defined as:

\[
\psi_i[k] = \begin{cases} 
  a, & k \leq k_{i_{\text{lo}}} \\
  \frac{k - k_{i_{\text{lo}}}}{k_{i_{\text{hi}}} - k_{i_{\text{lo}}}}, & k_{i_{\text{lo}}} \leq k \leq k_{i_{\text{hi}}} \\
  \frac{k_{i_{\text{hi}}}}{k_{i_{\text{hi}}} - k_{i_{\text{lo}}}} - k, & k_{i_{\text{lo}}} \leq k \leq k_{i_{\text{hi}}} \\
  0, & k \geq k_{i_{\text{hi}}}
\end{cases}
\]

Where \( i=1,\ldots,Q \), and \( Q \) denoted the number of filters in the bank; \( k \) denoted the spectral sample index in the \( M \)-points DFT, and \( k_{i_{\text{lo}}} \) were the spectral sample indexes denoting
equally spaced (on the logarithmic mel scale) boundary points of the filters. These boundary points were defined as:

\[
k_{b_j} = \left( \frac{M_s}{F_s} \right) \cdot f_{mel}^{-1} \left[ f_{mel}(f_{low}) + \frac{i\{f_{mel}(f_{high}) - f_{mel}(f_{low})\}}{Q + 1} \right]
\]  

(4-4)

Where \( F_s \) was the sampling frequency, \( f_{low} \) and \( f_{high} \) were the lowest and highest frequencies in Hz defining frequency range covered by the filter bank such that \( f_{low} = \frac{F_s}{M_s} \), \( f_{high} = \frac{F_s}{2} \). Finally, \( f_{mel}^{-1} \) was the inverse transformation operator of the mel scale, defined as:

\[
f_{mel}^{-1}(f_{mel}) = 700 \cdot \left[ \frac{f_{mel}}{10^{2595}} - 1 \right]
\]  

(4-5)

Following the above description, the filter bank was imposed on the spectrum calculated in Eq.(4-2) and the output samples \( e[i] \) were calculated from the bank of the mel-scaled band pass filters using:

\[
e[i] = \sum_{k=1}^{M_s} |y[k]|^2 \cdot \psi_i[k]
\]  

(4-6)

**Step 3:** The discrete cosine transform (DCT) was applied to the log filter bank energies \( \log(e[i]) \), and the mel frequency cepstral coefficients \( C_m \) \( m=0,1,\ldots,R-1 \) were calculated using:

\[
C_m = \sqrt{2} \sum_{l=0}^{Q-1} \log(e[i+1]) \cdot \cos\left( m \cdot \left( \frac{2l - 1}{2} \right) \cdot \frac{\pi}{Q} \right)
\]  

(4-7)

Where \( R \) was the desired number of MFCCs.

In the described here experiments, the number of MFCCs was determined experimentally as 12.
4.4 Estimation of the glottal waveform

During the process of speech production, the volume of the air flow generated by lung is passed through the glottis to form the glottal flow (Figure 4-1). The air passing through the glottis causes vibration of the vocal folds and generates a periodic sound wave called the glottal waveform. The glottal wave is then modulated by the elements of the vocal tract (filter).

As illustrated in Figure 4-4, the resulting speech can be therefore interpreted as a convolution of the glottal waveform $e(t)$ and the impulse response of the vocal tract filter. This indicates that the glottal waveform $e(t)$ can be extracted from the speech signal $s(t)$ through the process of deconvolution [174-179].

![Diagram of speech production process](image)

**Figure 4-4** Time and frequency domain representation of the speech production process. Speech samples $s[n]$ are represented in the time domain as a convolution of the glottal, vocal filter and lip radiation impulse responses. In the frequency domain the speech $z$-transform $S(z)$ is given as a product of the glottal, vocal filter and lip radiation transfer functions.

Unfortunately, the deconvolution operation belongs to one of the most challenging tasks in signal processing and the existing methods suffer from different types of
limitations. An exact representation of the glottal waveform is very difficult to achieve and only estimates can be calculated.

Several studies pointed to the correlation between the glottal features and the vocal expressions of emotional states. There is considerable evidence that the affect arousal is capable of modifying the dynamics of the air flow through the vocal folds and therefore produces changes in the glottal waveform characteristics. Cumming et al. [174-177] showed that when the speech was produced during a stressful situation, the glottal waveform was altered due to the excessive tension or lack of coordination in the laryngeal musculature. Different levels of stress were characterized by different patterns of these changes. Moore et al. [81, 82] analyzed the effectiveness of glottal features in speech based classification of clinical depression. The results indicated that glottal source features indicating the opening and closing instances and their ratio combination sets, showed good separation ability between depressed and non-depressed speakers. Ozdas et al., [125] also investigated the glottal flow spectrum as a possible cue for depression and near-term suicide risk showing the slope of the glottal flow spectrum provides good correct classification rates up to 85%.

The most often used methods for source-filter separation include the cepstral analysis and the linear predictive (LP) analysis [104]. The glottal waveform estimation method used in this thesis was based on the inverse filtering and linear predictive analysis.

Using the classical source-filter theory, the speech production process can be modeled as a series configuration of linear systems illustrated in Figure 4-4 [178, 179].

The parameter $e[n]$ in Figure 4-4 represents the excitation signal. For a voiced speech, $e[n]$ is represented as a quasi-periodic train of impulses, passing through a glottal model filter with impulse response $g[n]$ and transfer function $G(z)$ to generate the glottal velocity signal $u_g[n]$. The glottal flow velocity is then transformed by the vocal tract filter with impulse response $v[n]$ and transfer function $V(z)$. Finally the $u_v[n]$ is passed thought the lip radiation filter with impulse response $r[n]$ and transfer function $R(z)$ producing the speech samples $s[n]$. 
Based on the above process of speech production, the $z$-transform (or the frequency domain representation) of the glottal waveform was estimated as:

$$ U_o(z) = \frac{S(z)}{V(z)R(z)} \tag{4-8} $$

The lip radiation transfer function $R(z)$ in Eq.(4.8) was estimated as:

$$ R(z) = 1 - z^{-1} \tag{4-9} $$

The transfer function (inverse filter) $V(z)$ in Eq.(4-8), was modeled be the following all-pole function:

$$ V(z) = \frac{1}{1 + \sum_{i=1}^{K} c_i z^{-i}} \tag{4-10} $$

$V(z)$ was calculated using the VOICEBOX speech processing toolbox [146].

In the described here experiments the LP order (the inverse filter order) was experimentally set to $K=12$.

4.5 Glottal time and frequency domain features

![Figure 4-5 Schematic of the glottal waveform indicating the time intervals corresponding to the “open” and “closed” phases; $T$ denotes the glottal period.](image)

As illustrated in Figure 4-5, the glottal waveform is a quasi-periodic signal, where one cycle consists of two phases: the open phase (starting at $t_0$ and ending at $t_c$) and the closed
phase (staring at $t_c$ and ending at $t_0$). The combined time duration of the open and closed phases defines the time duration of the glottal period $T$.

Based on the glottal wave analysis, 14 time domain and 3 frequency domain glottal feature parameters were calculated. All calculations were performed using algorithms from the TTK Aparat glottal inverse filtering toolbox [180]. The time domain features included:

- $t_{\text{max}}$ and $t_{\text{min}}$ - the time instances corresponding to the maximum and minimum amplitude of the glottal time waveform;
- $t_{\text{dmax}}$ and $t_{\text{dmin}}$ - the time instances corresponding to the maximum and minimum amplitude of the glottal flow derivative;
- $t_{\text{c}}$ - the closing time of the glottis, estimated as the first positive zero-crossing after the minimum instant of the glottal flow derivative;
- $t_{\text{o1}}$ - the primary glottis opening time, measured as 10% of $A_{ac}$ (where $A_{ac}=A_{\text{max}}-A_{\text{min}}$) and $A_{\text{max}}$ and $A_{\text{min}}$ are the maximum and minimum amplitude of the glottal waveform respectively;
- $t_{\text{o2}}$ - the secondary glottis opening time, located at the largest local maximum of the second derivative of the glottal flow;
- $t_{\text{qo}}$ and $t_{\text{qc}}$ - the quasi-open and closing durations defined as a time period when the glottal flow is 50% above the minimum flow or 50% below the maximum flow respectively.
- $OQ1$ and $OQ2$ – the open quotients measuring the relative portion of the primary and secondary open phase compared to the whole cycle duration.
- $SQ1$ and $SQ2$ – the speed quotients measuring the ratio of the duration of open phase to the duration of the closed phase.
- $ClQ$ – the closing quotient measuring the ratio of the closing phase to the whole cycle duration.
- $AQ$ – the amplitude quotient measuring the ratio of the flow peak-to-peak amplitude to the minimum peak of the pulse derivative ($A_{d\text{min}}$).
- $NAQ$ – the normalized amplitude quotient measuring the ratio of the amplitude
quotient to the cycle duration.

- $QOO$ – the quasi-open quotient defined as a ratio of quasi-open phase to the cycle duration.

Since even slight non-linearity in the phase response of the recording equipment, could affect the quality of the glottal waveform estimation and subsequently have an effect on the above time domain parameters, three frequency domain parameters were also estimated. These parameters included:

- $dH1-dH2$ - measuring the difference between amplitudes of the first and second harmonic components of the glottal waveform.

- Harmonic richness factor (HRF) - measuring the ratio of the sum of the amplitudes of the higher harmonics to the amplitude of the first harmonic:  
  \[
  HRF = \left( \sum_{k \geq 2} H_k \right) / H_1 , \text{ where } H_k \text{ represents the amplitude of the } k\text{-th harmonic.}
  \]

- Parabolic spectral parameter – this parameter is based on fitting a parabolic polynomial to the low frequency part of the glottal waveform spectrum. It gives a single value which describes how the spectral decay of the glottal waveform behaves with respect to a theoretical limit corresponding to a maximum spectral decay. It was introduced to compensate for the possible detrimental effects of the fundamental frequency on $dH1-dH2$ and HRF.

### 4.6 Experiments and results

The stress and emotion classification tests were conducted using five types of features: fundamental frequency $F_0$, formants, MFCCs, glottal time domain parameters (GP-T) and glottal frequency domain parameters (GP-F). The modeling and classification was performed using two types of classifiers: GMM and KNN.

The speech data for stress classification included datasets from the Speech under Actual Stress domain of the SUSAS database described in Chapter 3 and representing three levels of stress: high, low and neutral. The speech data for emotion classification
CHAPTER 4. STRESS AND EMOTION CLASSIFICATION USING CLASSICAL FEATURES

included dataset from the ORI database described in Chapter 3 and representing five different emotions: neutral, anger, anxious, dysphoric and happy.

For each feature/classifier combination, the classification process was repeated 15 times, each time with different randomly selected training (80% of recordings) and testing (20% of recordings) data sets. An average percentage of the identification accuracy (APIA%) was then calculated over 15 repetitions using the formula given in Eq.(3-9).

4.6.1 Results of stress classification

Table 4-1 shows APIA% values obtained for stress classification using $F_0$, formants, MFCC, GP-T and GP-F features and two different classifiers GMM and KNN. It can be observed that both classifiers provide very similar classification rates. Figures 4-6 and 4-7 show the same results as Table 4-1 but in the bar form. Figure 4-6 compares the correct classification rates for different features and different data sets using the GMM classifier, whereas Figure 4-7 shows the same comparison but for the KNN classifier.

### Table 4-1 APIA% for stress classification using $F_0$, formants, MFCC, GP-T and GP-F features.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$F_0$</th>
<th>Formants</th>
<th>MFCC</th>
<th>GP-T</th>
<th>GP-F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
</tr>
<tr>
<td>ē vowel</td>
<td>57.88</td>
<td>51.21</td>
<td>61.21</td>
<td>67.88</td>
<td>75.15</td>
</tr>
<tr>
<td>ā vowel</td>
<td>52.18</td>
<td>46.55</td>
<td>74.36</td>
<td>79.09</td>
<td>68.55</td>
</tr>
<tr>
<td>single vowels</td>
<td>50.00</td>
<td>43.41</td>
<td>56.81</td>
<td>66.81</td>
<td>56.03</td>
</tr>
<tr>
<td>SUAS</td>
<td>49.97</td>
<td>43.71</td>
<td>52.39</td>
<td>61.54</td>
<td>58.87</td>
</tr>
</tbody>
</table>

Figure 4-6 APIA% for stress classification using $F_0$, formants, MFCC, GP-T and GP-F features combined with the GMM classifier.
CHAPTER 4. STRESS AND EMOTION CLASSIFICATION USING CLASSICAL FEATURES

Figure 4-7 APIA% for stress classification using $F_0$, formants, MFCC, GP-T and GP-F features combined with the KNN classifier.

For each feature/classifier combination the classification tests were repeated 15 times, each time using different randomly selected testing and training sets. Results from these 15 runs were analyzed statistically using the ANOVA method (from the SPSS package) to determine statistical significance of the differences in classification results provided by different feature/classifier combinations. The ANOVA TF statistics are presented in Table 4-2 with + signs indicating statistically significant differences ($p<0.005$) and – signs indicating statistically insignificant difference ($p>=0.005$).

Table 4-2 TF statistics using ANOVA for stress (SUSAS data sets) using the GMM classifier; + denotes statistically significant difference ($p<0.05$), - denotes statistically insignificant difference ($p>=0.05$).

<table>
<thead>
<tr>
<th></th>
<th>$F_0$</th>
<th>Formants</th>
<th>MFCC</th>
<th>GP-T</th>
<th>GP-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_0$</td>
<td>N/A</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Formants</td>
<td>+</td>
<td>N/A</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>MFCC</td>
<td>+</td>
<td>+</td>
<td>N/A</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>GP-T</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>N/A</td>
<td>+</td>
</tr>
<tr>
<td>GP-F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Finally, Table 4-3 shows an example of confusion table providing classification rates obtained for each stress level using the MFCC features and the GMM classifier.

Table 4-3 Confusion matrixes (%) for stress classification using MFCC features combined with the GMM classifier; H-high-level stress, L-low-level stress, N-neutral.

<table>
<thead>
<tr>
<th>Actual level of stress</th>
<th>Classified level of stress</th>
<th>ē vowel</th>
<th>ā vowel</th>
<th>Single vowels</th>
<th>SUAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H</td>
<td>70.77</td>
<td>23.85</td>
<td>5.38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>27.86</td>
<td>62.14</td>
<td>10.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>15.00</td>
<td>50.00</td>
<td>35.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>64.29</td>
<td>26.19</td>
<td>9.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>31.36</td>
<td>58.18</td>
<td>10.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>23.33</td>
<td>38.33</td>
<td>38.33</td>
<td></td>
</tr>
</tbody>
</table>

72
CHAPTER 4. STRESS AND EMOTION CLASSIFICATION USING CLASSICAL FEATURES

4.6.2 Results of emotion classification

The emotion classification results are presented in tables and graphs in the similar way to the results obtained for the stress classification. Following this pattern, Table 4-4 shows APIA% values obtained for emotion classification using F0, formants, MFCC, GP-T and GP-F features and two different classifiers GMM and KNN. It can be observed that both classifiers provide very similar classification rates. Figures 4-8 show the same results as Table 4-4 but in the bar form.

Table 4-4 APIA% for emotion classification using F0, formants, MFCC, GP-T and GP-F features.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>F0</th>
<th>Formants</th>
<th>MFCC</th>
<th>GP-T</th>
<th>GP-F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
</tr>
<tr>
<td>ORI</td>
<td>38.30</td>
<td>30.50</td>
<td>51.50</td>
<td>44.90</td>
<td>49.00</td>
</tr>
</tbody>
</table>

Figure 4-8 APIA% for emotion classification using F0, formants, MFCC, GP-T and GP-F features.

Like in the case of stress, for each feature/classifier combination the emotion classification tests were repeated 15 times, each time using different randomly selected testing and training sets. Results from these 15 runs were analyzed statistically using the ANOVA method (from the SPSS package) to determine statistical significance of the differences in classification results provided by different feature/classifier combinations. The ANOVA TF statistics are presented in Table 4-5 with + signs indicating statistically significant differences ($p<0.005$) and – signs indicating statistically insignificant differences ($p\geq0.005$).
Table 4-5 TF statistics using ANOVA for emotions (ORI data set) using the GMM classifier; + denotes statistically significant difference (p<0.05), - denotes statistically insignificant difference (p>=0.05).

<table>
<thead>
<tr>
<th></th>
<th>F₀</th>
<th>Formants</th>
<th>MFCC</th>
<th>GP-T</th>
<th>GP-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>F₀</td>
<td>N/A</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Formants</td>
<td>+</td>
<td>N/A</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>MFCC</td>
<td>+</td>
<td>-</td>
<td>N/A</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>GP-T</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>N/A</td>
<td>-</td>
</tr>
<tr>
<td>GP-F</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 4-6 shows an example of confusion table providing classification rates obtained for each of the five emotions using the GMM classifier and the MFCC features.

Table 4-6 Confusion matrixes (%) for emotion classification using MFCC features combined with the GMM classifier.

<table>
<thead>
<tr>
<th>Actual emotions</th>
<th>Classified emotions</th>
<th>Angry</th>
<th>Anxious</th>
<th>Dysphoric</th>
<th>Neutral</th>
<th>Happy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>44.00</td>
<td>11.50</td>
<td>10.50</td>
<td>10.50</td>
<td>23.50</td>
<td></td>
</tr>
<tr>
<td>Anxious</td>
<td>9.00</td>
<td>48.50</td>
<td>25.00</td>
<td>8.00</td>
<td>9.50</td>
<td></td>
</tr>
<tr>
<td>Dysphoric</td>
<td>2.00</td>
<td>14.50</td>
<td>49.00</td>
<td>28.00</td>
<td>6.50</td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>4.00</td>
<td>9.00</td>
<td>24.50</td>
<td>58.00</td>
<td>4.50</td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td>20.00</td>
<td>12.00</td>
<td>9.00</td>
<td>13.50</td>
<td>45.50</td>
<td></td>
</tr>
</tbody>
</table>

4.7 Conclusions and discussions

4.7.1 Stress classification using classical features

Table 4-1 as well as Figures 4-6 and 4-7 indicate, that in the case of stress classification, different linguistic levels of analysis (different data sets) performed differently depending on the type of the classifier and the type of features. Classification results based on words containing single vowels ē and ā provided generally higher rates than the classification based on words with mixed single vowels and the whole set of SUAS words. This could be explained by the fact that single vowels have stable patterns of the F₀, formants, MFCC, GP-T and GP-F features, some people can even recognize these vowels from their prosodic or spectral patterns by visual inspection. Words on the other hand, contain voiced elements grouped in different orders which increase the variability and text dependency of prosodic and glottal patterns, and it is therefore likely to decrease the classification rates.
Looking at the performance of different feature/classifier combinations presented in Table 4-1 and Figures 4-6&4-7, it appears that for both classifiers, the highest classification rates were provided by the MFCC parameters, the glottal time parameters (GP-T) and formants. The fundamental frequency $F_0$ and the glottal frequency parameters (GP-F) provided relatively poor performance.

Based on the statistical analysis presented in Table 4-2, the only feature pairs that did not provide statistically significant differences in the classification rates (using GMM classifier) were: $F_0$ versus GP-F, and MFCC versus GP-T. This means that the differences between the best performing parameters (MFCC, GP-T, formants), and the low performance parameters ($F_0$, GP-F) were statistically significant.

The example of confusion between classes illustrated in Table 4-3 shows that in the case of MFCC/GMM combination there was a relatively high confusion level due to the fact that neutral speech is often miss-classified as low level stress. Generally there was a good differentiation between high level stress and low level stress as well as between high level stress and neutral speech.

The best stress classification result for the SUAS data was 65% and it was obtained using the MFCC features and the KNN classifier (Table 4-1).

### 4.7.2 Emotion classification using classical features

Analyzing the performance of different feature/classifier combinations presented in Table 4-4 and Figures 4-8, it appears that in the case of emotion classification for both classifiers, the highest classification rates were provided by formants and the MFCC parameters. The fundamental frequency $F_0$ and the glottal time (GP-T) and frequency (GP-F) provided relatively poor performance.

Based on the statistical analysis presented in Table 4-5 (for the GMM classifier), there was no statistically significant difference between the formants and MFCC performance and both types of features provided relatively high classification rates. There were also no statistically significant differences between pair combinations of $F_0$ and the glottal and
the glottal features (GP-T and GP-F) which provided relatively poor performance. However the differences between the pair combinations including the high performing parameters (formants and MFCC) and the low performing parameters ($F_0$, GP-T and GP-F) were statistically significant.

Inspecting the example of confusion between classes illustrated in Table 4-6, it can be observed that in the case of MFCC/GMM combination there was a relatively high confusion level due to the happy emotion being classified as angry. Otherwise a good differentiation between emotional classes can be observed.

The best emotion classification result was 51% and it was obtained for the ORI data using formants and the KNN classifier (Table 4-4).
This Chapter introduces a number of new feature extraction methods for an automatic detection of stress and emotion in speech. The proposed feature extraction methods use features derived from the speech spectrograms, wavelet packets (WP) and the empirical mode decomposition (EMD). The proposed features are tested using the same speech data sets and classifiers as those used with classical features tested in Chapter 4.

5.1 New feature extraction methods based on speech spectrograms

5.1.1 Advantages of using speech spectrograms

A two-dimensional magnitude spectrogram is a graphical display of the magnitude of the time-varying spectral characteristics of speech. It can be used to calculate numerous parameters such as energy, fundamental frequency (F₀), formants and timing. These parameters are the acoustic features of speech most often used in automatic stress and emotion recognition systems [91]. The majority of these systems analyze each parameter separately, and then combine them into a set of feature vectors. The approach presented here aimed to capture all of these characteristics at once, and preserve the important underlying dependencies between different parameters through analysis of speech spectrograms.

The speech spectrograms were previously not applied to the stress and emotion recognition problem, however other closely related applications have been reported. Kleinschmidt et al. [181, 182] applied a 2D Gabor filter bank to mel-spectrograms. The resulting outputs of the Gabor filters were concatenated into one-dimensional vectors and used as features in the speech recognition experiments. Chih et al. [183] applied a similar method to the process speech discrimination and enhancement. In recent studies Ezzat et
al. [184-187] described a spectro-temporal Gabor filter bank and used it to analyze localized patches of spectrograms, which showed advantages over one-dimensional features in word recognition. Meyer applied Gabor-shaped localized spectro-temporal features to successfully enhance automatic speech recognition performance [188].

5.1.2 Calculation of speech spectrograms

Speech spectrograms were calculated using short-time Fourier analysis applied to 256-point frames of voiced speech with 50% overlap. The global maximum of the absolute magnitude was calculated for each spectrogram, and the absolute magnitude level at 50dB below the maximum value was chosen as the minimum and set to 0dB. All absolute magnitudes below the minimum level were also set to 0dB, and all absolute magnitudes between the minimum and maximum levels were mapped into the range of 0dB-50dB. The 50dB value was determined experimentally as providing the best classification results.

5.1.3 Inspection of spectrograms for speech under stress and emotion

![Figure 5-1](image.png)

**Figure 5-1** Examples of spectrograms for vowel ā in the word “break” pronounced by the same parson; soft speech (a), neutral speech (b), speech under low level stress (c) and speech under high level stress (d).
CHAPTER 5. NEW FEATURE EXTRACTION METHODS BASED ON SPEECH SPECTROGRAMS, WAVELET PACKET AND EMPIRICAL MODE DECOMPOSITION

Figure 5-1 shows examples of the spectrograms for vowel ā pronounced by the same person for soft speech and under different stress levels. It can be observed that, with increasing level of stress, the spectrograms revealed increasing formant energy in the higher frequency bands, as well as clearly increasing pitch for high level stress. Other acoustic information, such as the formants also vary under different levels of stress. These observations indicate that the spectrograms contain important characteristics that can be used to differentiate between different levels of stress.

Figure 5-2 Examples of spectrograms for voiced parts of sentences pronounced with different emotions. Note that each spectrogram corresponds to a different sentence.

Figure 5-2 shows examples of speech spectrograms for sentences from the ORI data base recorded under different emotions. It should be noted that the examples in Figure 5-2 were calculated using different utterances for each emotion. Since the ORI data represented natural speech, it was technically difficult to find the same utterances expressed with different emotions. Despite of this limitation, Figure 5-2 provides a number of important observations.

Firstly, it shows that different emotions are characterized by amplitude gradients and distributions of energy across frequencies. Secondly, the spectral energy decreases with frequency, however the rate of this decrease differs across different emotions.
Using the above findings, a number of new feature extractions was proposed and tested as described in the following Sections.

### 5.1.4 New features based on speech spectrograms

The spectrograms were used to calculate the following types of new features for stress and emotion recognition:

- Features extracted from sub-bands of speech spectrograms without anisotropic filtering (SS-CB-AE, SS-BARK-AE, SS-ERB-AE),
- Features extracted from sub-bands of speech spectrograms with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE, SS-AF-ERB-AE),
- Features extracted from speech spectrograms and combined with a single log-Gabor filter and an optimal feature selection (SS-LGF-OFS),
- Features extracted from speech spectrograms and combined with a log-Gabor filter-bank, averaging and an optimal feature selection (SS-ALGF-OFS),
- Feature generation using spectrogram patches and log-Gabor filters (SS-SP-ALGF-OFS),
- Features extracted from speech spectrograms and sigma-pi units (SS-sigma-pi).

### 5.1.5 Features extracted from speech sub-bands of spectrograms using different auditory scales (SS-CB-AE, SS-BARK-AE, SS-ERB-AE)

(A) Frequency bands

**Critical bands**

The process of speech perception by the human auditory system shows high sensitivity to sounds occurring within specific frequency bands called the critical bands. The critical bands were introduced by Fletcher and Munson in the 1940s, who referred to the frequency bandwidth of the then loosely defined auditory filter [189]. Critical bands were determined using simple listening tests. The listeners were presented with a pure tone submersed in white noise of a limited bandwidth. The amplitude of the tone was gradually decreased and the level was recorded when the listener could no longer hear it.
The bandwidth of the noise was then reduced and the test was repeated. It was found, that
the level where the listener was unable to hear the tone remained the same until the
bandwidth of the noise was reduced to a critical width. Once the bandwidth of the noise
was within this critical width, the listener’s ability to hear the tone increased. It was
concluded, that the auditory system works like a bank of filters. The widths of the filters
are constrained to the critical points on either side of the tone and any noise outside this
region is ignored. The critical points collected by Fletcher and Munson can be used as a
scale to describe human hearing. The bands, defined by the region between the critical
points, represent the bandwidths of the filters in the ear’s psychological filter-bank.
Fletcher and Munson named their bands, the critical bands.

Table 5-1 provides a list of lower and upper boundaries of the critical bands in Hz
within the speech bandwidth ranging from 0 to 4 KHz. The widths of the critical bands
increase logarithmically with frequency and the centre frequencies \( B_c \) in Hz are equally
distant on the log scale.

Zhou’s [91], demonstrated that the extraction of characteristic features based on critical
bands was an important factor increasing the correct classification rates in an automatic
stress classification.

Since the work of Fletcher and Munson other types of auditory scales have been
developed. Two of the most popular scales are the bark scale and the equivalent
rectangular bandwidth scale.

**Bark scale**

The Bark scale is a psychoacoustical scale proposed by Eberhard Zwicker in 1961
[189]. The Bark scale represents critical bands rates given by a parameter \( z \) in Barks.

For a given frequency \( f \) in Hz, the corresponding values of \( z \) in Barks can be calculated
as follows:

\[
z = \left[ 26.81 / (1 + 1960 / f) \right] - 0.53
\]  (5- 1)
The inverse operation is given as:

\[ f = \frac{1960}{[26.81/(z + 0.53) - 1]} \]  

(5-2)

The Bark scale bands \( B_{\text{Bark}} \) in Hz can be then calculated for different values of \( z \) as follows:

\[ B_{\text{Bark}} = \frac{52548}{z^2 - 52.56z + 690.39} \]  

(5-3)

The lower and upper edges of the Bark scale bands within the range 0 to 4kHz are listed in Table 5-1.

Table 5-1 Analysis frequency bands (critical bands, Bark scale and ERB scale).

<table>
<thead>
<tr>
<th>No</th>
<th>Critical Bands [Hz]</th>
<th>Bark Scale [Hz]</th>
<th>ERB Scale [Hz]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
<td>Lower</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>200</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>300</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>400</td>
<td>155</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>510</td>
<td>250</td>
</tr>
<tr>
<td>5</td>
<td>510</td>
<td>630</td>
<td>345</td>
</tr>
<tr>
<td>6</td>
<td>630</td>
<td>770</td>
<td>450</td>
</tr>
<tr>
<td>7</td>
<td>770</td>
<td>920</td>
<td>565</td>
</tr>
<tr>
<td>8</td>
<td>920</td>
<td>1080</td>
<td>700</td>
</tr>
<tr>
<td>9</td>
<td>1080</td>
<td>1270</td>
<td>830</td>
</tr>
<tr>
<td>10</td>
<td>1270</td>
<td>1480</td>
<td>990</td>
</tr>
<tr>
<td>11</td>
<td>1480</td>
<td>1720</td>
<td>1170</td>
</tr>
<tr>
<td>12</td>
<td>1720</td>
<td>2000</td>
<td>1365</td>
</tr>
<tr>
<td>13</td>
<td>2000</td>
<td>2320</td>
<td>1590</td>
</tr>
<tr>
<td>14</td>
<td>2320</td>
<td>2700</td>
<td>1850</td>
</tr>
<tr>
<td>15</td>
<td>2700</td>
<td>3150</td>
<td>2145</td>
</tr>
<tr>
<td>16</td>
<td>3150*</td>
<td>3700</td>
<td>2505</td>
</tr>
<tr>
<td>17</td>
<td>3700*</td>
<td>4310*</td>
<td>2910</td>
</tr>
<tr>
<td>18</td>
<td>4310</td>
<td>5110</td>
<td>3410</td>
</tr>
<tr>
<td>19</td>
<td>5110</td>
<td>5910</td>
<td>4110</td>
</tr>
<tr>
<td>20</td>
<td>5910</td>
<td>6710</td>
<td>4810</td>
</tr>
<tr>
<td>21</td>
<td>6710</td>
<td>7510</td>
<td>5510</td>
</tr>
<tr>
<td>22</td>
<td>7510</td>
<td>8310</td>
<td>6210</td>
</tr>
<tr>
<td>23</td>
<td>8310</td>
<td>9110</td>
<td>6910</td>
</tr>
<tr>
<td>24</td>
<td>9110</td>
<td>9910</td>
<td>7610</td>
</tr>
<tr>
<td>25</td>
<td>9910</td>
<td>10710</td>
<td>8310</td>
</tr>
<tr>
<td>26</td>
<td>10710</td>
<td>11510</td>
<td>9010</td>
</tr>
<tr>
<td>27</td>
<td>11510</td>
<td>12310</td>
<td>9710</td>
</tr>
</tbody>
</table>

**Equivalent Rectangular Bandwidth (ERB) scale**

Since the hearing system performs a temporal analysis that contributes to frequency resolution for low frequencies, auditory frequency resolution cannot be fully represented on the basis of \( z \) alone.
It has been postulated that, the auditory frequency resolution is better described by the equivalent rectangular bandwidth (ERB) [190, 191]. The equivalent rectangular bandwidth is a measure of auditory frequency bands, which approximates the auditory system as a bank of rectangular band-pass filters. The ERB bandwidth values in Hz for a given center frequency \( f \) in Hz can be calculated as:

\[
B_{ERB} = 6.23 \cdot 10^{-6} f^2 + 9.339 \cdot 10^{-2} f + 28.52
\]  

(5-4)

The lower and upper edges of the ERB scale bands within the range 0 to 4kHz are listed in Table 5-1.

**B Feature generation (SS-CB-AE, SS-BARK-AE and SS-ERB-AE)**

The 2D spectrograms of energy spectral density (squared magnitudes) were divided into sub-bands based on three different auditory scales: critical bands, Bark scale, and ERB scale. For each sub-band a single value of the average energy \( \hat{E}_i \) (\( i=1,\ldots,N \)) was calculated using:

\[
\hat{E}_i = \frac{1}{N_f N_t} \sum_{t=1}^{N_t} \sum_{f=1}^{N_f} s(t,f)
\]  

(5-5)

Where \( s(t,f) \) are the spectrogram values (squared magnitudes) at the time coordinates \( t \) and frequency coordinates \( f \), \( N_f \) is the total number of frequency coordinates, \( N_t \) is the total number of time coordinates, and \( N \) is the total number of frequency bands (\( N=16 \) for critical bands, \( N=17 \) for the Bark scale and \( N=27 \) for the ERB scale [192]).

---

**Figure 5-3** Features generation from the auditory frequency bands of spectrograms using different auditory scales (SS-CB-AE, SS-BARK-AE and SS-ERB-AE).
The resulting feature values were then concatenated into 1D vectors, and passed to the GMM and KNN classifiers for modeling and classification. The flow chart of the feature extraction process is illustrated in Figure 5-3.

5.1.6 Features extracted from speech spectrograms with anisotropic filtering and different auditory scales (SS-AF-CB-AE, SS-AF-BARK-AE, SS-AF-ERB-AE)

(A) Anisotropic Diffusion and Nonlinear Filtering

The anisotropic diffusion filtering of images, also called Perona–Malik diffusion [193], is a technique aimed at reducing image noise without removing significant parts of the image content. Anisotropic filtering was previously successfully used in biomedical image processing to reduce noise and enhance contrast in specific regions of images. Gerig et al. [194] applied an anisotropic filtering technique to the 2-D and 3-D spin echo and gradient echo magnetic resonance (MR) data. Ding et al. [195] tested anisotropic smoothing on in-vivo diffusion tensor data for noise reduction.

For $\Omega \in R^2$ denoting a subset of the plane and $s(\cdot, \tau): \Omega \rightarrow R^2$ being a spectrogram (image), the implementation of anisotropic diffusion filtering was defined as a sum of functions $\Phi(\cdot)$ called directional flows:

$$\frac{\partial s}{\partial \tau} = \Phi_{East} - \Phi_{West} + \Phi_{North} - \Phi_{South}$$  \hspace{1cm} (5-6)

The directional flow functions are defined as:

$$\Phi_{East (West)} = \frac{1}{\Delta x^2} \left( c(x \pm \Delta x, y, \tau) \star (s(x + \Delta x, y, \tau) - s(x, y, \tau)) \right)$$  \hspace{1cm} (5-7)

$$\Phi_{North (South)} = \frac{1}{\Delta y^2} \left( c(x, y \pm \Delta y, \tau) \star (s(x, y + \Delta y, \tau) - s(x, y, \tau)) \right)$$  \hspace{1cm} (5-8)

The parameter $c(x,y,\tau)$ is the diffusion coefficient controlling the rate of diffusion and is usually chosen as the following function of the image gradient $\|\nabla s\|$ so as to preserve the edges in the image:
\[ c(\|\nabla \mathbf{s}\|) = \exp\left(-\frac{\|\nabla \mathbf{s}\|}{\kappa}\right) \] (5-9)

Where the constant \( \kappa \) controls the sensitivity to edges and is usually chosen experimentally.

Based on the 2-D spectrogram topography, the signal flows were firstly calculated between neighboring pixels. The pixels’ intensities were then updated as the following local sum of the flow contributions:

\[
s(\tau + \Delta \tau) \approx s(\tau) + \Delta \tau \ast \frac{\partial s(\tau)}{\partial \tau} = s(\tau) + \Delta \tau \ast \left( \Phi_{\text{East}} - \Phi_{\text{West}} + \Phi_{\text{North}} - \Phi_{\text{South}} \right) \] (5-10)

(B) Feature Generation (SS-AF-CB-AE, SS-AF-BARK-AE and SS-AF-ERB-AE)

Compared to the feature generation approach described in Section 5.1.5 in this approach an anisotropic diffusion filtering was used to perform the enhancement of spectrograms before dividing them into sub-bands.

After the anisotropic diffusion filtering, the average energy for each sub-band was calculated and the results were concatenated into 1D feature vectors (Figure 5-4).

Figure 5-4 Feature extraction method using auditory frequency bands of anisotropic filtering spectrograms (SS-AF-CB-AE, SS-AF-BARK-AE and SS-AF-ERB-AE).

5.1.7 Experiments and results for features SS-CB-AE, SS-BARK-AE, SS-ERB-AE, SS-AF-CB-AE, SS-AF-BARK-AE and SS-AF-ERB-AE)

The stress and emotion classification tests were conducted using six types of features extracted from speech spectrograms: SS-CB-AE, SS-BARK-AE, SS-ERB-AE, SS-AF-CB-AE, SS-AF-BARK-AE and SS-AF-ERB-AE. The modeling and classification was performed using two types of classifiers: GMM and KNN.
The speech data for stress classification included datasets from the Speech under Actual Stress domain of the SUSAS database described in Chapter 3 and representing three levels of stress: high, low and neutral. The speech data for emotion classification included dataset from the ORI database described in Chapter 3 and representing five different emotions: neutral, anger, anxious, dysphoric, and happy.

For each feature/classifier combination, the classification process was repeated 15 times, each time with different randomly selected training (80%) and testing (20%) data sets. An average percentage of the identification accuracy (APIA%) was then calculated over 15 repetitions using the formula given in Eq.(3-9).

**(A) Results of stress classification**

Table 5-2 shows APIA% values obtained for stress classification using spectrogram features without (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) and with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE and SS-AF-ERB-AE) and two different classifiers GMM and KNN. Figures 5-5 and 5-6 show the same results as Table 5-2 but in a bar form for the GMM and KNN classifies respectively.

It can be observed in Tables 5-2 that in both cases without and with anisotropic filtering, the ERB based features (SS-ERB-AE and SS-AF-ERB-AE) provided the best average classification accuracy.

**Table 5-2** APIA% for stress classification using spectrogram features without anisotropic filtering (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) and with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE, SS-AF-ERB-AE).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Spectrogram features without anisotropic filtering</th>
<th>Spectrogram features with anisotropic filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>ē vowels</td>
<td>75.76</td>
<td>75.15</td>
</tr>
<tr>
<td>à vowels</td>
<td>66.55</td>
<td>73.45</td>
</tr>
<tr>
<td>Single vowel</td>
<td>58.41</td>
<td>66.98</td>
</tr>
<tr>
<td>SUAS</td>
<td>59.28</td>
<td>67.61</td>
</tr>
</tbody>
</table>
CHAPTER 5. NEW FEATURE EXTRACTION METHODS BASED ON SPEECH SPECTROGRAMS, WAVELET PACKET AND EMPIRICAL MODE DECOMPOSITION

Figure 5-5 APIA% for stress classification using spectrogram features without anisotropic filtering (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) and with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE and SS-AF-ERB-AE) combined with the GMM classifier; NF-without anisotropic filtering, AF-with anisotropic filtering.

Figure 5-6 APIA% for stress classification using spectrogram features without anisotropic filtering (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) and with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE and SS-AF-ERB-AE) combined with the KNN classifier; NF-without anisotropic filtering, AF-with anisotropic filtering.

Table 5-3 and Table 5-4 show examples of the confusion matrix (percentage of correct classification and misclassification within each class) using the GMM classifier and the best performing features (SS-ERB-AE in Table 5-3 and the SS-AF-ERB-AE in Table 5-4).
Table 5-3 Confusion table for stress classification using SS-ERB-AE feature combined with the GMM classifier; H-high level stress, L-low level stress, N-neutral

<table>
<thead>
<tr>
<th>Actual level of stress</th>
<th>Classified level of stress</th>
<th>Spectrogram features without anisotropic filtering</th>
<th>SUAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>e vowel</td>
<td>a vowel</td>
<td>single vowels</td>
</tr>
<tr>
<td>H</td>
<td>90.00</td>
<td>9.23</td>
<td>0.77</td>
</tr>
<tr>
<td>L</td>
<td>14.29</td>
<td>84.29</td>
<td>1.43</td>
</tr>
<tr>
<td>N</td>
<td>13.33</td>
<td>28.33</td>
<td>58.33</td>
</tr>
</tbody>
</table>

Table 5-4 Confusion table for stress classification using SS-AF-ERB-AE feature combined with the GMM classifier; H-high level stress, L-low level stress, N-neutral

<table>
<thead>
<tr>
<th>Actual level of stress</th>
<th>Classified level of stress</th>
<th>Spectrogram features with anisotropic filtering</th>
<th>SUAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>e vowel</td>
<td>a vowel</td>
<td>single vowels</td>
</tr>
<tr>
<td>H</td>
<td>73.85</td>
<td>25.38</td>
<td>0.77</td>
</tr>
<tr>
<td>L</td>
<td>16.43</td>
<td>77.14</td>
<td>6.43</td>
</tr>
<tr>
<td>N</td>
<td>1.67</td>
<td>13.33</td>
<td>85.00</td>
</tr>
</tbody>
</table>

(B) Results of emotion classification

The emotion classification results are presented in a similar way to the stress classification results, thus Table 5-5 shows APIA% values obtained for stress classification using spectrogram features without (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) and with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE and SS-AF-ERB-AE) and two different classifiers GMM and KNN. Figure 5-7 shows the same results but in a form of bar graph.

It can be observed in Tables 5-5 that in both cases without and with anisotropic filtering, the ERB based features (SS-ERB-AE and SS-ERB-AE) provided the best average classification accuracy.

Table 5-5 APIA% for emotion classification using spectrogram features without anisotropic filtering (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) and with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE, SS-AF-ERB-AE).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Spectrogram features without anisotropic filtering</th>
<th>Spectrogram features with anisotropic filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM KNN GMM KNN GMM KNN</td>
<td>GMM KNN GMM KNN GMM KNN</td>
</tr>
<tr>
<td>ē vowels</td>
<td>48.80 47.33 51.40 49.83 53.40 52.50</td>
<td>43.40 43.83 45.60 45.83 51.50 48.50</td>
</tr>
</tbody>
</table>
CHAPTER 5. NEW FEATURE EXTRACTION METHODS BASED ON SPEECH SPECTROGRAMS, WAVELET PACKET AND EMPIRICAL MODE DECOMPOSITION

Figure 5-7 APIA% for emotion classification using spectrogram features without anisotropic filtering (SS-CB-AE, SS-BARK-AE and SS-ERB-AE) and with anisotropic filtering (SS-AF-CB-AE, SS-AF-BARK-AE, SS-AF-ERB-AE).

Table 5-6 shows examples of the confusion matrix using the GMM classifier and the best performing features SS-ERB-AE and SS-AF-ERB-AE.

Table 5-6 Confusion table for the emotion classification using SS-ERB-AE and SS-AF-ERB-AE features combined with the GMM classifier; AE-Actual emotions, Ag-anger, Ax anxious, Dy-dysphoric, Ne-neutral, Ha-happy.

<table>
<thead>
<tr>
<th>AE</th>
<th>Classified emotions</th>
<th>SS-ERB-AE feature</th>
<th>SS-AF-ERB-AE feature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>GMM</td>
<td>KNN</td>
</tr>
<tr>
<td>Ag</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ax</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ne</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ha</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.1.8 Discussion of the results for SS-AF-ERB-AE, SS-AF-CB-AE, SS-AF-BARK-AE, SS-CB-AE, SS-BARK-AE and SS-ERB-AE features

(A) Effects of frequency sub-division

The classification results for stress and emotions showed that the best classification rates in both cases with and without anisotropic filtering were obtained while dividing the spectrograms into ERB bands and then calculating average energy for each sub-band.
Subdivision into Bark scale band produced lower results, and the worse results were obtained while using critical bands.

Due to differences in bandwidth definition, the ERB bands are narrower than the classical critical bands at all frequencies. These results indicate that characteristic features are located both at high and low frequencies and fine division across the whole bandwidth of speech signal is essential in the process of stress and emotion recognition. Unlike CB or Bark scales, the ERB-scale does not only represent the tonotopic structure of human auditory system but also incorporates a temporal analysis that contributes to frequency resolution for low frequencies.

(B) Effects of an anisotropic filtering

The effects of an anisotropic filtering of spectrograms on the stress and emotion classification rates in speech were tested. The anisotropic filtering was applied to spectrograms prior to the calculation of the average spectral energy for the frequency sub-bands. The results presented led to the following conclusions:

1) In the case of stress detection using words with single wolves, a clear improvement of classification results due to the anisotropic filtering was observed.

2) For stress detection based on a closed set of single words, the anisotropic filtering provided only a very small improvement of the classification results.

3) In the case of long utterances consisting of words from an open set used in emotion detection, a clear decrease of the classification results was observed after the anisotropic filtering of spectrograms.

These results can be explained by the existence of characteristic patterns in spectrograms of single vowels. These patterns are so well defined that many linguists can recognize vowels simply by visual inspection of their spectrograms.

In the case of words from closed or open sets which contain different mixtures of voiced components (vowels or/and consonants) such patterns cannot be in general detected.
unless a prior detection of vowels and consonants is made. In the 2D spectrograms of vowels, formants form characteristic pattern lines proceeding along the time axis with certain amount of obliquity (detraction from the East-West direction). It is therefore possible that the use of more than 2 directions (East-West and North-South) during the calculation of image flow parameters could further improve the resolution of formant trajectories.

(C) Stress versus emotions

Generally, higher classification rates obtained for SUSAS data (stress levels) than for ORI data (types of emotions) can be explained by the fact that the SUSAS corpora contained speech produced under highly stressful conditions (rollercoaster and pilot’s cockpit) when the arousal levels can be expected to be very high. The ORI data on the other hand, was recorded during a typical family discussion, when emotions are usually expressed with low or mild levels of emotional arousal.

(D) Confusion between different stress levels and between different types of emotions

The examples of confusion tables (Table 5-3 and 5-4) for the SUSAS database (classification of stress levels) show certain amount of confusion or misclassification between the low level stress and the neutral speech. The high level stress on the other hand is usually well distinguished from the neutral speech and from the low level stress.

A possible explanation can be derived from the visual inspection of Figure 5-1, which shows examples of spectrograms for the same word “break” (containing a single vowel ā) pronounced by the same parson under soft speech, neutral speech, speech under low level stress and speech under high level stress. It can be seen in this example that, the high level stress shows a distinct pattern with relatively high energy levels at high frequency range. This pattern is highly distinguishable from the patterns corresponding to the low level stress and neutral speech spectrograms. At the same time, it can be observed, that the spectral patterns for the low level stress and neutral speech are very similar, and both show relatively low energy content at high frequencies.
The examples of the confusion tables (Table 5-6) for different emotions show that anger was most often confused with happiness and the anxious emotion was most often confused with the dysphoric emotional state. In these cases, an inspection of spectral patterns could not lead to clear conclusions because of difficulties of finding the same sentences expressed with different emotions in a spontaneous speech. However, spectral patterns for different sentences showed in Figure 5-2 can provide some clues as to why anger is confused with happiness and anxious feeling is confused with dysphoric feeling. This is again based on the similarity of spectral energy distribution corresponding to these emotions.

Figure 5-2 also indicates that a negative emotion such as anger which is often associated with a high level of stress produces high energies at the high frequency range, whereas more passive emotional states such as dysphoria and anxiety produce relatively low energy content at high frequencies.

(E) The best performing feature/classifier combinations

The best stress classification result for stress was obtained when using SS-AF-ERB-AE/GMM, whereas the best classification of emotions was given by SS-ERB-AE.

5.1.9 Stress and emotion recognition using log-Gabor filter analysis of spectrograms and optimal feature selection (SS-LGF-OFS, SS-ALGF-OFS and SS-SP-ALGF-OFS)

(A) Log-Gabor filters

Gabor filters are commonly recognized [197] as one of the best choices for obtaining features in image classification. They offer an excellent simultaneous localization of spatial and frequency information. However, the maximum bandwidth of a Gabor filter is limited to approximately one octave and Gabor filters are not optimal if one is seeking broad spectral information with maximal spatial localization.

As an alternative to the Gabor filters the log-Gabor filters were proposed by Field [198-200]. Log-Gabor filters can be constructed with arbitrary bandwidth and the bandwidth
can be optimized to produce a filter with minimal spatial extent. The log-Gabor filters have Gaussian transfer functions when viewed on the logarithmic frequency scale, whereas the Gabor filters have Gaussian transfer functions when viewed on the linear frequency scale. It was therefore postulated that the log-Gabor functions having extended tails at the high frequency ends should be able to encode natural images more efficiently by better representing the higher frequency components. The transfer functions of log-Gabor filters are compatible with the human visual system, which has cell responses that are symmetric on the log frequency scale. Furthermore, a log-Gabor Filter always has a zero DC component and therefore, the filter bandwidth can be optimized to produce a filter with minimal spatial extent.

The log-Gabor filters in the frequency domain can be defined in polar coordinates by the transfer function $G(r, \theta)$ constructed as the following product:

$$G(r, \theta) = G_{\text{radial}}(r) \cdot G_{\text{angular}}(r)$$

(5-11)

Where $G_{\text{radial}}(r)$ is the frequency response of the radial component given as:

$$G_{\text{radial}}(r) = \exp\left(-\log(r/f_0)^2/2\sigma_{\text{radial}}^2\right)$$

(5-12)

And $G_{\text{angular}}(r)$ represents the frequency response of the angular filter component, given as:

$$G_{\text{angular}}(r) = \exp\left(- (\theta - \theta_0)^2/2\sigma_\theta^2 \right)$$

(5-13)

In Eq.(5-11) – Eq.(5-13), (r, \theta) are the polar coordinates, $f_0$ represents the central filter frequency, $\theta_0$ is the orientation angle, $\sigma_r$ and $\sigma_\theta$ represent the scale bandwidth and angular bandwidth respectively.

In the described here experiments, the number of different wavelengths r (scales) for the filter bank was set to $N_r=2$, and for each wavelength of the filter the number of different orientations $\theta$ was set to $N_\theta=6$. This produced a bank of 12 log-Gabor filters $\{G_1, G_2, \ldots, G_{12}\}$ with each filter representing different scale and orientation.
The log-Gabor feature representation $|S(t,f)|_{m,n}$ of a magnitude spectrogram $s(t,f)$ was calculated as a convolution operation performed separately for the real and imaginary part of the log-Gabor filters:

\[
\text{Re}(S(t,f))_{m,n} = s(t,f) \ast \text{Re}(G(r_m, \theta_n))
\]

\[
\text{Im}(S(t,f))_{m,n} = s(t,f) \ast \text{Im}(G(r_m, \theta_n))
\]

Where (t,f) represent the time and frequency coordinates of a spectrogram, and $m=1,\ldots,N_r=2$ and $n=1,\ldots,N_\theta=6$. This was followed by the magnitude calculation for the filter bank outputs,

\[
|S(t,f)|_{m,n} = \sqrt{[\text{Re}(S(t,f))_{m,n}]^2 + [\text{Im}(S(t,f))_{m,n}]^2}
\]

(B) Averaging outputs of the log-Gabor filters

For each spectrogram, the log-Gabor filter bank outputs were averaged to produce a single output array:

\[
|\hat{S}(t,f)| = \frac{1}{N_r N_\theta} \sum_{m=1}^{N_r} \sum_{n=1}^{N_\theta} |S(t,f)|_{m,n}
\]

The averaged arrays were then converted to 1D vectors via a row-by-row concatenation.

(C) Optimal feature selection using mutual information (MI) criteria

The total set of $N_F$ feature vectors was reduced to a small sub-set of $N_s < N_F$ vectors selected using the mutual information feature selection algorithm described in [196][201, 202]. The mutual information represents a measure of information found commonly in two random variables $X$ and $Y$, and it is given as:

\[
I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
\]
Where \( p(x) \) is the probability density function (pdf), defined as \( p(x) = \Pr\{X=x\} \), and \( p(x,y) \) is the joint pdf defined as \( p(x,y) = \Pr(X=x \text{ and } Y=y) \).

Given an initial set \( F \) with \( N_F \) feature vectors and a set \( C \) of all output classes (\( C=\{1,2,3\} \) for SUSAS data and \( C=\{1,2,3,4,5\} \) for ORI data, the aim was to find an optimal subset \( S \) with \( N_S < N_F \) feature vectors. Starting from the empty set, the best available feature vectors were added, one by one to the selected feature set, until the size of the set reached the desired value of \( N_S \). The sub-set \( S \) of feature vectors was selected through simultaneous maximization of the mutual information between the selected feature vectors in \( S \) and the class labels \( C \), and minimization of the mutual information between the selected feature vectors within \( S \). As a result an optimal sub-set \( S \) of mutually independent and highly representative feature vectors was obtained [203, 204].

Given the full set size of \( N_F=513 \) (using SUSAS data) we tested the classification process using optimal sub-set sizes of \( N_s=10, 20, 30, 40, 50, 60 \) and 70. The results showed that \( N_s=10 \) gives the best compromise between the classification accuracy and the data reduction rate. The same value of \( N_s=10 \) was then used in experiments with the ORI data.

**D) Features extracted from speech spectrograms and combined with a single log-Gabor filter and an optimal feature selection (SS-LGF-OFS)**

The spectrograms of voiced part of speech signals were calculated and passed thought a bank of 12 log-Gabor filters with 2 different scales and 6 different orientations (Figure 5-8). For each filter the magnitudes of the filter outputs were then passed through an optimal feature selection algorithm based on mutual information (MI) criteria, and used in modeling and classification of stress and emotions. The results allowed determining which scales and orientations of log-Gabor filters give the best classification scores.

**Figure 5-8** Features generation using single log-Gabor filter (SS-LGF-OFS).
(E) Features extracted from speech spectrograms and combined with a log-Gabor filter-bank, averaging and an optimal feature selection (SS-ALGF-OFS)

The spectrograms of voiced part of the speech samples were calculated and passed thought a bank of 12 log-Gabor filters. The 12 outputs from the log-Gabor filters were averaged, passed through the MI feature selection, and sent to the classifiers (Figure 5-9).

(F) Feature generation using spectrogram patches and log-Gabor filters (SS-SP-ALGF-OFS)

The spectrograms of voiced part of speech signals were calculated, and three patches (or time-frequency regions) were extracted from each spectrogram. The first patch included frequencies from 0.05Hz to 355.2Hz, the second patch, from 355.2Hz to 1290.7Hz, and the third patch, from 1290.7Hz to 3754.5Hz. The patches were determined experimentally to cover the low, middle and high frequency ranges. Each patch was passed through 12 log-Gabor filters. The outputs were averaged and passed thought the MI feature selection algorithm, then the GMM and KNN were used to classify different stress levels or different types of emotions (Figure 5-10).
5.1.10 Experiments and results for SS-LGF-OFS, SS-ALGF-OFS and SS-SP-ALGF-OFS

Stress and emotion classification tests were conducted using three types of features extracted from speech spectrograms: SS-LGF-OFS, SS-ALGF-OFS and SS-SP-ALGF-OFS. The modeling and classification was performed using two types of classifiers: GMM and KNN.

The speech data for stress classification included datasets from the Speech under Actual Stress domain of the SUSAS database described in Chapter 3 and representing three levels of stress: high, low and neutral. The speech data for emotion classification included dataset from the ORI database described in Chapter 3 and representing five different emotions: neutral, anger, anxious, dysphoric, and happy.

For each feature/classifier combination, the classification process was repeated 15 times, each time with different randomly selected training (80%) and testing (20%) data sets. An average percentage of the identification accuracy (APIA%) was then calculated over 15 repetitions using the formula given in Eq.(3-9).

(A) Results of stress classification (SS-LGF-OFS, SS-ALGF-OFS and SS-SP-ALGF-OFS)

Table 5-7 shows APIA% values obtained for stress classification using speech spectrograms combined with a single log-Gabor filter (SS-LGF-OFS) and the GMM and KNN classifier. Each of the 12 log-Gabor filters in Tables 5-7 is characterized by its orientation (n) and the scale factor (m). Figure 5-11 and Figure 5-12 show the same results as Table 5-7 but in a bar form for the GMM and KNN classifiers respectively.

Table 5-7 APIA% for stress classification for each of the 12 log-Gabor filters (SS-LGF-OFS) using the GMM and KNN classifiers.

<table>
<thead>
<tr>
<th>Scale (m)</th>
<th>Orientation (n)</th>
<th>Orientation (degree)</th>
<th>APIA%</th>
<th>Single vowels</th>
<th>SUAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>à vowels</td>
<td>ã vowels</td>
<td>GMM</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0°</td>
<td>56.77</td>
<td>50.71</td>
<td>47.03</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>18°</td>
<td>61.21</td>
<td>59.19</td>
<td>54.06</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>36°</td>
<td>61.01</td>
<td>57.98</td>
<td>59.39</td>
</tr>
</tbody>
</table>
CHAPTER 5. NEW FEATURE EXTRACTION METHODS BASED ON SPEECH SPECTROGRAMS, WAVELET PACKET AND EMPIRICAL MODE DECOMPOSITION

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>54°</th>
<th>63.84</th>
<th>68.28</th>
<th>58.67</th>
<th>58.67</th>
<th>59.80</th>
<th>60.95</th>
<th>60.10</th>
<th>61.89</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>72°</td>
<td>57.78</td>
<td>59.60</td>
<td>55.03</td>
<td>57.82</td>
<td>61.06</td>
<td>60.40</td>
<td>59.56</td>
<td>59.43</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>90°</td>
<td>60.00</td>
<td>55.56</td>
<td>54.55</td>
<td>52.24</td>
<td>54.89</td>
<td>50.78</td>
<td>57.55</td>
<td>52.08</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0°</td>
<td>45.86</td>
<td>41.62</td>
<td>45.94</td>
<td>43.39</td>
<td>48.07</td>
<td>41.75</td>
<td>48.20</td>
<td>42.58</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>18°</td>
<td>51.72</td>
<td>49.09</td>
<td>48.85</td>
<td>41.45</td>
<td>49.25</td>
<td>44.97</td>
<td>49.41</td>
<td>46.44</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>36°</td>
<td>60.61</td>
<td>63.23</td>
<td>57.70</td>
<td>56.97</td>
<td>64.91</td>
<td>63.62</td>
<td>62.39</td>
<td>63.29</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>54°</td>
<td>67.27</td>
<td>71.31</td>
<td>64.36</td>
<td>67.15</td>
<td>62.79</td>
<td>65.29</td>
<td>61.61</td>
<td>64.07</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>72°</td>
<td>64.04</td>
<td>59.80</td>
<td>62.42</td>
<td>56.73</td>
<td>61.03</td>
<td>58.36</td>
<td>60.63</td>
<td>56.56</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>90°</td>
<td>52.73</td>
<td>51.52</td>
<td>47.88</td>
<td>42.79</td>
<td>51.38</td>
<td>47.87</td>
<td>49.22</td>
<td>45.58</td>
</tr>
</tbody>
</table>

Figure 5-11 APIA% for stress classification for each of the 12 log-Gabor filters (SS-LGF-OFS) and the GMM classifier; S-scale, O-Orientation.

Figure 5-12 APIA% for stress classification for each of the 12 log-Gabor filters (SS-LGF-OFS) and the KNN classifier; S-scale, O-Orientation.
Table 5-8 presents the APIA% values when using the SS-ALGF-OFS and SS-SP-ALGF-OFS features for the GMM and KNN classifiers. Figure 5-13 shows the same results but in a form of bar graph.

**Table 5-8** APIA% for stress classification using SS-ALGF-OFS and SS-SP-ALGF-OFS features.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SS-ALGF-OFS</th>
<th>SS-SP-ALGF-OFS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
</tr>
<tr>
<td>ē vowel</td>
<td>77.58</td>
<td>75.76</td>
</tr>
<tr>
<td>ā vowel</td>
<td>79.03</td>
<td>72.91</td>
</tr>
<tr>
<td>single vowels</td>
<td>73.76</td>
<td>64.74</td>
</tr>
<tr>
<td>SUAS</td>
<td>64.7</td>
<td>64.25</td>
</tr>
</tbody>
</table>

![Figure 5-13](image)

**Figure 5-13** APIA% for stress classification using SS-ALGF-OFS and SS-SP-ALGF-OFS features.

(B) **Results of emotion classification (SS-LGF-OFS, SS-ALGF-OFS and SS-SP-ALGF-OFS)**

The emotion classification results are presented in the similar way as the results for stress classification. Table 5-9 shows APIA% values obtained for emotion classification using SS-LGF-OFS features and the GMM and KNN classifiers. Figure 5-14 shows the same results but in a bar graph.
Table 5-9 APIA% for emotion classification for each of the 12 log-Gabor filters (SS-LGF-OFS) using the GMM and KNN classifiers.

<table>
<thead>
<tr>
<th>Scale (m)</th>
<th>Orientation (n)</th>
<th>Orientation (degree)</th>
<th>GMM</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0°</td>
<td>43.88</td>
<td>42.17</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>18°</td>
<td>43.88</td>
<td>38.83</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>36°</td>
<td>42.50</td>
<td>38.83</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>54°</td>
<td>43.63</td>
<td>38.83</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>72°</td>
<td>42.75</td>
<td>38.83</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>90°</td>
<td>45.00</td>
<td>43.83</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0°</td>
<td>39.63</td>
<td>38.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>18°</td>
<td>41.13</td>
<td>42.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>36°</td>
<td>37.50</td>
<td>39.67</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>54°</td>
<td>45.37</td>
<td>38.33</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>72°</td>
<td>44.50</td>
<td>41.17</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>90°</td>
<td>40.75</td>
<td>40.67</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5-14 APIA% for emotion classification for each of the 12 log-Gabor filters (SS-LGF-OFS) using the GMM and KNN classifiers; S-scale, O-orientation.

Table 5-10 shows APIA% values obtained when using the SS-ALGF-OFS and SS-SP-ALGF-OFS features and the GMM and KNN classifiers. Figure 5-15 shows the same results but in a bar graph.

Table 5-10 APIA% for emotion classification using SS-ALGF-OFS and SS-SP-ALGF-OFS features.
(C) Discussions of results for stress and emotion classification using SS-LGF-OFS, SS-ALGF-OFS and SS-SP-ALGF-OFS

For both stress and emotion classification, Tables 5-7 to Table 5-10 show that both classifiers provided very similar trends and the differences between the APIA% values for these two classifiers were insignificant.

The classification results based on single filters (SS-LGF-OFS) are relatively low, ranging from 45% to 67% for the stress data (SUSAS) and from 37% to 45% for the emotion data (ORI) while using GMM classifier, and ranging from 42% to 71% for the stress data and from 38% to 43% for the emotion data while using KNN classifier.

Figure 5-16 shows an example of the classification accuracy (APIA%) versus the log-Gabor filter orientation when using two scales m=1 and m=2 for words with ā vowels and the GMM classifier. The best performing filters in this case appeared to be filters characterized by the scale m=2 and orientation n=4 (54°) and n=5 (72°). However this result cannot be easily generalized to all types of voiced speech; different combinations of vowels and voiced consonants may show high correlation along different orientations.

To overcome this problem of finding an optimal orientation of a log-Gabor filter, the outputs from all of the 12 log-Gabor filters (each with different orientation) were

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SS-ALGF-OFS</th>
<th>SS-SP-ALGF-OFS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
</tr>
<tr>
<td>ORI</td>
<td>39.60</td>
<td>40.67</td>
</tr>
</tbody>
</table>

Figure 5-15 APIA% for emotion classification using SS-ALGF-OFS and SS-SP-ALGF-OFS features.
averaged and a small sub-set of optimal features was selected using the mutual information criteria in the SS-ALGF-OFS approach.

![Figure 5-16 APIA (%) versus filter orientation (degree) for two scales m=1 (blue line) and m=2 (pink line) for words with only ā vowel and the GMM classifier.](image)

**Figure 5-16** APIA (%) versus filter orientation (degree) for two scales m=1 (blue line) and m=2 (pink line) for words with only ā vowel and the GMM classifier.

The classification results for SS-ALGF-OFS (Tables 5-8 (stress) and 5-10 (emotions)) are clearly showing an improvement over a single log-Gabor filter method SS-LGF-OFS. For the averaged log-Gabor filters (SS-ALGF-OFS), the results are close to those obtained in our previous approaches with the ERB frequency bands of spectrograms (SS-AF-ERB-AE and SS-AF-ERB-AE) and range from 64% to 79% for the SUSAS data and 39% for the ORI data when using the GMM classifier, and from 64% to 75% for the SUSAS data and 40% for the ORI data when using the KNN classifier.

The classification results (SS-SP-ALGF-OFS, Tables 5-8 (stress) and 5-10 (emotions)) obtained while applying 12 log-Gabor filters to 3 spectrogram patches representing low, medium and high frequency bands are slightly lower than for the averaged 12 log-Gabor filters and range from 61% to 73% for the SUSAS data and 40% for the ORI data while using the GMM classifier, and from 54% to 74% for the SUSAS data and 39% for the ORI data while using the KNN classifier.

When comparing the discussed here 3 methods: SS-LGF-OFS, SS-ALGF-OFS and SS-SP-ALGF-OFS, the best overall performance for both classifiers was obtained for the SS-ALGF-OFS (averaged 12 log-Gabor filters) feature. This can be attributed to the fact that the averaging process attenuates the uncorrelated spectral features and enhances the
correlated features on the time frequency plane. The SS-LGF-OFS on the other hand chooses an arbitrary direction along the time-frequency plane which may not be optimal from the perspective of stress or emotion classification. Similarly, the SS-SP-ALGF-OFS method uses an arbitrary bandwidth sub-division into 3 patches which also may not be optimal.

5.1.11 Features extracted from speech spectrograms using sigma-pi units (SS-sigma-pi)

(A) Feature extraction method

The voiced speech signals were used to calculate speech spectrograms as the primary features. As illustrated in Figure 5-17, the primarily features were then used to compute secondary features using the window based sigma-pi neurons described in [205, 206].

The sigma-pi neurons were calculated from the spectrograms based on three alternative bandwidth sub-divisions into the following frequency bands: critical bands (CB), Bark scale, and Equivalent Rectangular Bandwidth (ERB).

(B) Sigma-pi unit

The sigma-pi neuron unit is known as the second-order feature in neural network theory with applications in speech recognition [205, 206]. In 1990’s, Gramms [207, 208] proposed a method using sigma-pi cells for word recognition as the second order feature to detect the beginning, end and velocity of the tonal changes. The Sigma-pi units were later used by Kleinschmidt [209] in combination with spectrograms for the task of recognition of isolated words. The results showed that sigma-pi cells provide features suitable for automatic speech recognition. In Kleinschmidt’s later work [210], it was demonstrated that sigma-pi cells achieve a good performance as secondary features for the sub-band signal-to-noise ratio (SNR) estimation. The presented algorithm increased
the computational effort relative to a short-term SNR system, since the introduction of secondary features made the feature vectors sparser, allowing for easier integration over all input values and better linear classification.

A sigma-pi neutron was defined as the multiplicative element of the weighted output of two or more units over all input values.

At a given time instance \( t \), each sigma-pi neuron calculated the following product, which was derived using information from two time-frequency windows as illustrated in Figure 5-18.

\[
S(t_1, t_2, f_1, f_2, t) = w_1 \cdot w_2 
\]  
(5-19)

The values of \( w_1 \) and \( w_2 \) for the two windows were defined in the following way:

\[
w_1 = s(t + t_1, f_1) 
\]  
(5-20)

\[
w_2 = \frac{1}{\Delta t_2 \Delta f_2} \sum_{f_1} \sum_{f_2} s(t + t_2 + t', f_2 + f') 
\]  
(5-21)

Where \( s(t,f) \) represents the primary feature (spectrogram) values, \( \Delta t_2 \) and \( \Delta f_2 \) represent the time and frequency extensions for the second analysis windows. The values of \( f' \) and \( t' \) were defined to meet the following criteria:

\[-\Delta f_2 / 2 \leq f' \leq \Delta f_2 / 2, -\Delta t_2 / 2 \leq t' \leq \Delta t_2 / 2 \]  
(5-22)

The time extensions \( \Delta t_2 \) was kept at a constant value of 3. The frequency extension values \( \Delta f_2 \) varied according to the type of the alternative analysis bandwidth of critical bands (CB), Bark scale (Bark) or ERB scale (ERB):

\[
\Delta f_2 = \frac{1}{3} \text{Bandwidth(CB, Bark or ERB)} 
\]  
(5-23)

The analysis bands corresponding to the CB, Bark and ERB scales are listed in Table 5-1.
Eq. (5-20) and (5-21) show that, the $w_1$ value for the first window uses only one primary element, whereas, $w_2$ uses more than one element, which could solve the time invariance problem [208].

\[
\begin{bmatrix}
S_1 \\
S_2 \\
\vdots \\
S_N \\
\end{bmatrix}
\]

Where $N$ is the number of frequency bands.

**Figure 5-18** Structure of a Sigma-pi neuron.  **Figure 5-19** Generation of $S_i$ values over speech frames.

The sigma-pi neuron values were integrated over all speech frames within each frequency band producing an integrated $S_i$ value for each frequency band $i$. This process is illustrated in Figures 5-18 and 5-19.

Finally, a secondary feature vector $\Sigma$ was made by concatenation of $S_i$ values for all frequency bands:

\[
\Sigma = \begin{bmatrix}
S_1 \\
S_2 \\
\vdots \\
S_N \\
\end{bmatrix}
\]  \hspace{1cm} (5-24)
5.1.12 Experiments and results for sigma-pi units (SS-sigma-pi)

The stress and emotion classification tests were conducted using the SS-sigma-pi features. The modeling and classification was performed using two types of classifiers: GMM and KNN.

Stress and emotion classification tests were conducted using three types of features extracted from speech spectrograms: SS-LGF-OFS, SS-ALGF-OFS and SS-SP-ALGF-OFS. The modeling and classification was performed using two types of classifiers: GMM and KNN.

The speech data for stress classification included datasets from the Speech under Actual Stress domain of the SUSAS database described in Chapter 3 and representing three levels of stress: high, low and neutral. The speech data for emotion classification included dataset from the ORI database described in Chapter 3 and representing five different emotions: neutral, anger, anxious, dysphoric, and happy.

For each feature/classifier combination, the classification process was repeated 15 times, each time with different randomly selected training (80%) and testing (20%) data sets. An average percentage of the identification accuracy (APIA%) was then calculated over 15 repetitions using the formula given in Eq.(3-9).

(A) Results of stress classification using SS-sigma-pi

Table 5-11 shows the classification accuracy for stress calculated as APIA% using the SS-sigma-pi features and two classifiers GMM and KNN. The time distance parameter in Table 5-11 denotes the time distance between two windows in each of the sigma-pi neurons; two time distances were used: 32ms and 48 ms. The results were calculated using three different bandwidth subdivisions: into critical bands (CB), into Bark bands (Bark) and into ERB bands (ERB). Figures 5-20 shows the same results as Table 5-11 but in a bar form using time distance of 48 ms.

<table>
<thead>
<tr>
<th>Frequency bands</th>
<th>Time</th>
<th>ź vowel</th>
<th>ź vowel</th>
<th>single vowels</th>
<th>SUAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5-11 APIA% for stress classification using the SS-sigma-pi features and two classifiers: GMM and KNN.
CHAPTER 5. NEW FEATURE EXTRACTION METHODS BASED ON SPEECH SPECTROGRAMS, WAVELET PACKET AND EMPIRICAL MODE DECOMPOSITION

<table>
<thead>
<tr>
<th>Frequency bands</th>
<th>Time distance</th>
<th>GMM</th>
<th>KNN</th>
<th>GMM</th>
<th>KNN</th>
<th>GMM</th>
<th>KNN</th>
<th>GMM</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical bands</td>
<td>32 ms</td>
<td>57.02</td>
<td>66.76</td>
<td>59.45</td>
<td>63.18</td>
<td>57.62</td>
<td>59.45</td>
<td>59.45</td>
<td>61.95</td>
</tr>
<tr>
<td></td>
<td>48 ms</td>
<td>58.51</td>
<td>65.87</td>
<td>57.68</td>
<td>62.45</td>
<td>55.34</td>
<td>58.12</td>
<td>55.57</td>
<td>62.01</td>
</tr>
<tr>
<td>Bark scale</td>
<td>32 ms</td>
<td>70.25</td>
<td>70.35</td>
<td>67.22</td>
<td>69.42</td>
<td>65.52</td>
<td>66.77</td>
<td>62.63</td>
<td>70.09</td>
</tr>
<tr>
<td></td>
<td>48 ms</td>
<td>64.38</td>
<td>68.16</td>
<td>65.63</td>
<td>68.20</td>
<td>64.26</td>
<td>64.66</td>
<td>65.00</td>
<td>68.08</td>
</tr>
<tr>
<td>ERB scale</td>
<td>32 ms</td>
<td>72.93</td>
<td>72.24</td>
<td>71.56</td>
<td>73.76</td>
<td>70.15</td>
<td>70.12</td>
<td>67.36</td>
<td>72.96</td>
</tr>
<tr>
<td></td>
<td>48 ms</td>
<td>66.27</td>
<td>69.95</td>
<td>68.81</td>
<td>70.70</td>
<td>67.84</td>
<td>69.30</td>
<td>66.10</td>
<td>71.92</td>
</tr>
</tbody>
</table>

**Figure 5-20** APIA % for stress classification using the SS-sigma-pi features (48ms time distance) within CB, Bark and ERB bands and the GMM and KNN classifier.

(B) Results of emotion classification using SS-sigma-pi

The emotion classification results are presented in the similar way as the results for the stress classification. Table 5-12 shows the classification accuracy for emotions calculated as APIA% using the SS-sigma-pi features and two classifiers GMM and KNN. The time distance parameter in Table 5-12 denotes the time distance between two windows in each of the sigma-pi neurons; two time distances were used: 32ms and 48 ms. The results were calculated using three different bandwidth subdivisions: CB, Bark and ERB bands. Figures 5-21 shows the same results as Table 5-12 but in a bar form using time distance of 48 ms.

**Table 5-12** APIA% for emotion classification using the SS-sigma-pi features and two classifiers: GMM and KNN.

<table>
<thead>
<tr>
<th>Frequency bands</th>
<th>Time distance</th>
<th>GMM</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical bands</td>
<td>32 ms</td>
<td>40.50</td>
<td>42.60</td>
</tr>
<tr>
<td></td>
<td>48 ms</td>
<td>41.83</td>
<td>43.50</td>
</tr>
</tbody>
</table>
CHAPTER 5. NEW FEATURE EXTRACTION METHODS BASED ON SPEECH SPECTROGRAMS, WAVELET PACKET AND EMPIRICAL MODE DECOMPOSITION

<table>
<thead>
<tr>
<th></th>
<th>Bark scale</th>
<th>ERB scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time distance</td>
<td>32 ms</td>
<td>44.00</td>
</tr>
<tr>
<td></td>
<td>48 ms</td>
<td>42.00</td>
</tr>
<tr>
<td>Time distance</td>
<td>32 ms</td>
<td>46.50</td>
</tr>
<tr>
<td></td>
<td>48 ms</td>
<td>46.80</td>
</tr>
</tbody>
</table>

**Figure 5-21** APIA % for emotion classification using the SS-sigma-pi features (48ms time distance) within CB, Bark and ERB bands and the GMM and KNN classifiers.

(C) Discussion of stress and emotion classification results for SS-sigma-pi features

The results showed that both classifiers GMM and KNN yielded very similar correct classification rates for both stress and emotion.

For the stress data, there were also very small differences in the range of results obtained for four different data sets (57.02%-72.93% for vowel ē dataset, 57.68%-73.76% for vowel ā dataset, 55.34%-70.15% for mixed vowels dataset and 53.14%-72.96% for the SUAS dataset). This could possibly indicate that the stress conditions have very small effect on the shape of formants which define vowels, and therefore the proposed method can be applied using features extracted from voiced speech in general without looking at specific types of vowels.

In the case of emotion classification, the SS-sigma-pi features achieved the correct classification rates up to 46.80% for the ERB bands and GMM classifier using time distance between sigma-pi windows of 48 ms, and up to 48.8% for the ERB bands and KNN classifier using time distance between sigma-pi windows of 32 ms.
CHAPTER 5. NEW FEATURE EXTRACTION METHODS BASED ON SPEECH SPECTROGRAMS, WAVELET PACKET AND EMPIRICAL MODE DECOMPOSITION

The choice of the analytical frequency bands for the formation of feature had clearly the most significant effect on the achieved classification rates, as illustrated in Table 5-11 (for stress), and Table 5-12 (for emotions), the best classification results were achieved when using the ERB frequency band analysis. The second best was the Bark scale analysis and the lowest range of classification rates was obtained when using the critical band analysis. The ERB scale provides narrower frequency bands and higher frequency resolution at low frequencies when compare to the CB and Bark methods. The high performance of the ERB scale could therefore indicate that speech produced under stress and different emotions may have characteristics which have an effect on temporal analysis performed by the human auditory system while differentiating between low frequency sounds. One can speculate that similar effects are experienced by an automatic recognition system.

It was also observed that, the time distance between two windows in each sigma-pi neuron had an effect on the achieved classification accuracy. The sigma-pi cells with time distance of 32ms generally achieved higher accuracy rates than the sigma-pi features with 48ms time distance. It could indicate that acoustical characteristic for different stress levels and different types of emotions are hidden within fine pitch durations, displayed as the time distance in the speech spectrograms.

5.2 Features extracted from the wavelet packet arrays combined with a log-Gabor filter bank, averaging and an optimal feature selection (WP-ALGF-OFS)

5.2.1 WP-ALGF-OFS feature extraction process

Section 5.1 described different methods of feature extraction from the speech spectrograms which provide the image-like 2D time-frequency representation of speech characteristics based on the short-time Fourier transform.

This Section proposes another feature extraction method based on the 2D time-frequency representation of speech characteristics provided by the wavelet analysis.
A general flowchart of the proposed feature extraction method (WP-ALGF-OFS) is illustrated in Figure 5-22.

The wavelet packet array of coefficients was calculated for each frame of the voiced speech signal, and passed through a bank of 12 log-Gabor filters with two different scales (m=1 and m=2) and six different orientations (0°, 180°, 360°, 540°, 720°, 900°). The averaged outputs of the 12 log-Gabor filters were concatenated into 1D vectors and passed through the optimal feature selection algorithm based on the mutual information criteria (described in Section 5.1.9 (C)) to reduce the dimension of feature vectors to 10.

The modeling of stress and emotion classes and the classification process were performed using the GMM and the KNN methods.

**5.2.2 Calculation of the Wavelet Packet coefficients arrays**

The wavelet packet method offers a modified form of the discrete wavelet transform, where the signal is passed iteratively through twice the number of filters used in the DWT. Each level of decomposition was calculated by passing the previous level approximation and detail coefficients through the high and the low pass filters (see Figure 5-23). The WP analysis was expected to provide both, the low and the high frequency stress and emotion cues.
Figure 5-23 Wavelet Packet (WP) two-levels decomposition; Gi denotes the low pass filters, Hi denotes the high pass filters, $\downarrow 2$ denote dyadic division by a factor of 2.

The WP coefficients were calculated by decomposing the speech frames up to the 4th level of the wavelet packet’s tree with the db2 mother wavelet [211-213]. The frequency ranges corresponding to each of the terminal nodes are listed in Table 5-13.

<table>
<thead>
<tr>
<th>No.</th>
<th>Lower</th>
<th>Upper</th>
<th>No.</th>
<th>Lower</th>
<th>Upper</th>
<th>No.</th>
<th>Lower</th>
<th>Upper</th>
<th>No.</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>250</td>
<td>5</td>
<td>1000</td>
<td>1250</td>
<td>9</td>
<td>2000</td>
<td>2250</td>
<td>13</td>
<td>3000</td>
<td>3250</td>
</tr>
<tr>
<td>2</td>
<td>250</td>
<td>500</td>
<td>6</td>
<td>1250</td>
<td>1500</td>
<td>10</td>
<td>2250</td>
<td>2500</td>
<td>14</td>
<td>3250</td>
<td>3500</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>750</td>
<td>7</td>
<td>1500</td>
<td>1750</td>
<td>11</td>
<td>2500</td>
<td>2750</td>
<td>15</td>
<td>3500</td>
<td>3750</td>
</tr>
<tr>
<td>4</td>
<td>750</td>
<td>1000</td>
<td>8</td>
<td>1750</td>
<td>2000</td>
<td>12</td>
<td>2750</td>
<td>3000</td>
<td>16</td>
<td>3750</td>
<td>4000</td>
</tr>
</tbody>
</table>

The values of the WP coefficients provided 2D time frequency signal representations. Examples of the WP arrays are illustrated in Figure 5-24 and 5-25.
CHAPTER 5. NEW FEATURE EXTRACTION METHODS BASED ON SPEECH SPECTROGRAMS, WAVELET PACKET AND EMPirical MODE DECOMPOSITION

Figure 5-24 Examples of wavelet packet arrays for the vowel ā in the word “break” pronounced under (a) high-level stress speech, (b) low-level stress speech and (c) neutral speech. Horizontal axis-time, vertical axis-frequency; the color indicates the coefficient’s values: purple-high values of coefficients, blue-low values of coefficients.

Figure 5-25 Examples of wavelet packet arrays for utterances pronounced with the following emotions: (a) angry, (b) anxious, (c) dysphoric, (d) neutral and (e) happy. Note that each emotion is represented by a different utterance. Horizontal axis-time, vertical axis-frequency; the color
indicates the coefficient’s values: purple-high values of coefficients, blue-low values of coefficients.

Figure 5-24 shows examples of the 2D wavelet packet arrays of vowel ā in the word “break” pronounced under (a) high-level stress speech, (b) low-level stress speech and (c) neutral speech. Figure 5-25 shows examples of wavelet packet arrays for utterances pronounced with different emotions: (a) angry, (b) anxious, (c) dysphoric, (d) neutral and (e) happy. Note that each emotion in Figure 5-25 is represented by a different utterance, due to the technical difficult to find same sentences with different emotions.

It should be noted that the examples in Figure 5-25 were calculated using different utterances for each emotion. Since the ORI data represented natural speech it was not practical to search for identical utterances pronounced with different emotions. Despite of this limitation, Figure 5-24 and 5-25 provide a number of important observations.

Firstly, Figures 5-24 and 5-25 show that different stress levels and different emotions are characterized by different distribution patterns of the values of the WP coefficients across the time-frequency planes. Secondly, the spectral energy decreases with frequency, however the rate of this decrease differs across emotions and across different stress levels.

In particular Figure 5-24 shows that an increasing level of stress leads to an increase of WP coefficient’s values within the higher frequency bands. Similarly, it can be observed in Figure 5-25 that the anger shows higher WP coefficient’s values at higher frequency bands compare to other emotions, whereas the dysphoric emotion shows high coefficient’s values in a large span at the low frequency bands. The anxious, happy and neutral emotions show similar patterns of medium-valued coefficient’s evenly distributed across frequencies.

Based on the above findings, it was expected that the log-Gabor analysis of 2D WP coefficients representations could provide patterns characterizing different stress levels and emotions.
5.2.3 Experiments and results for the WP-ALGF-OFS features

The stress and emotion classification tests were conducted using the WP-ALGF-OFS feature and two types of classifiers: GMM and KNN.

The speech data for stress classification included datasets from the Speech under Actual Stress domain of the SUSAS database described in Chapter 3 and representing three levels of stress: high, low and neutral. The speech data for emotion classification included dataset from the ORI database described in Chapter 3 and representing five different emotions: neutral, anger, anxious, dysphoric, and happy.

For each feature/classifier combination, the classification process was repeated 15 times, each time with different randomly selected training (80%) and testing (20%) data sets. An average percentage of the identification accuracy (APIA%) was then calculated over 15 repetitions using the formula given in Eq.(3-9).

(A) Results of stress classification using WP-ALGF-OFS features

Table 5-14 shows APIA% values obtained for stress classification using WP-ALGF-OFS features and two classifiers GMM and KNN. Figure 5-26 shows the same results as the Table 5-14 and presents them in a form of a bar-graph.

Table 5-14 APIA% for stress classification using WP-ALGF-OFS features and two classifiers GMM and KNN.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GMM</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ē vowel</td>
<td>84.85</td>
<td>77.37</td>
</tr>
<tr>
<td>ā vowel</td>
<td>81.27</td>
<td>80.85</td>
</tr>
<tr>
<td>single vowels</td>
<td>76.72</td>
<td>76.72</td>
</tr>
<tr>
<td>SUAS</td>
<td>76.35</td>
<td>76.81</td>
</tr>
</tbody>
</table>
(B) Results of emotion classification using WP-ALGF-OFS features

Table 5-15 shows APIA% values obtained for emotion classification using WP-ALGF-OFS features and two classifiers GMM and KNN.

Table 5-15 APIA% for emotion classification using WP-ALGF-OFS features and two classifiers GMM and KNN.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GMM</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORI</td>
<td>45.50</td>
<td>40.00</td>
</tr>
</tbody>
</table>

(C) Discussion of results for WP-ALGF-OFS features

The wavelet packet method combined with the log-Gabor filters showed particularly promising results in the process of automatic stress and emotion classification in speech. Our results showed significantly lower classification rates for the ORI data base when compared with the data obtained from the SUSAS sets. This can be attributed to the different environments in which these two data bases were recorded. The SUSAS data base was generated for the purpose of research on stress and emotion detection, and contained speech recordings made during a rollercoaster ride when a very strong stress or emotion expression can be expected. The ORI data on the other hand contains spontaneously expressed emotions during typical family based conversation when, the
emotional expressions are not expected to be as strong as in the situations captured by the SUSAS data.

Figure 5-27 shows a comparison of APIA% values for WP-ALGF-OFS features and two previously introduced features: SS-ALGF-OFS and SS-ERB-AE, which achieved highest classification accuracy among the spectrogram based extraction features. The WP-ALGF-OFS features outperformed SS-ALGF-OFS and SS-ERB-AE in the case of stress recognition, however in the case of emotion recognition the performance of WP-ALGF-OFS was lower than the performance of SS-ALGF-OFS and SS-ERB-AE. It can be also observed in Figure 5-27 that in all approaches, the highest classification accuracy was achieved while using single vowels, which is not surprising since vowels are distinguished by characteristic time-frequency patterns. It is possible that the results for the ORI data could be improved if instead of voiced speech detection, automatic detection of particular vowels is used and the features are then extracted from time-frequency arrays representing these vowels.

![Figure 5-27](image.png)

**Figure 5-27** A comparison of APIA% for stress and emotion classification using SS-ALGF-OFS, SS-ERB-AE and WP-ALGF-OFS features.
5.3 Features derived from the Empirical Mode Decomposition (EMD) combined with calculation of an averaged Renyi’s entropy (EMD-AER)

5.3.1 Empirical Mode Decomposition (EMD)

As suggested in [91, 214], the airflow in the vocal tract is separated into different tracts, each with its own energy. These tracts include the main air flow through the vocal folds as well as additional vortex-flows generated due to specific emotional states. This model was further supported in [215] proposing to represent the speech signal as a sum of linearly separable components. Following these ideas, this study proposes a feature extraction method (Figure 5-28) that directly analyzes the time waveform of the speech signal into separate components called the intrinsic mode functions (IMF) [216, 217]. For each IMF channel the Renyi entropy [218] was calculated and then averaged over all channels. The resulting average value of the Renyi entropy was expected to be sensitive to changes in the spectral energy distribution resulting from stress and emotions.

![Figure 5-28](image)

Figure 5-28 Feature extraction using EMD and Renyi entropy (ER).

The EMD analysis was performed on a frame-by-frame basis. Speech frames of length $N=256$ samples with 50% overlap were used. The 256 sample window corresponded to a time length of 32 ms. Within this time duration, speech spectral characteristics could be considered stationary. It also ensured that a few glottal cycles were included within each frame. To calculate the IMFs, all the local minima and maxima of the speech signal $x[k]$ ($k=1,\ldots,N$) within a given frame were identified. The local maxima and minima were connected by cubic spline interpolation curves to generate upper and lower envelopes, respectively. The mean contour $m[k]$ was then calculated between the lower and upper envelopes, and used to generate the first approximation $h_1[k]$ of the first empirical mode.
component, which was defined as the difference between the data samples \( x[k] \) and the \( m_1[k] \) contour:

\[
h_1[k] = x[k] - m_1[k]
\]  
(5-25)

The \( h_1[k] \) values were then iteratively refined during the sifting process. In the second sifting iteration the \( h_1[n] \) was replaced by \( h_{11}[n] \) given as:

\[
h_{11}[k] = h_1[k] - m_{11}[k]
\]  
(5-26)

Where \( m_{11}[k] \) is an average contour between the upper and lower envelopes of \( h_1[k] \). After repeating the sifting process \( S_1 \) times, the first intrinsic mode function \( IMF_1[k] \) was given as:

\[
IMF_1[k] = h_{1S_1}[k] = h_{1S_1-1}[k] - m_{1S_1-1}[k]
\]  
(5-27)

The first IMF \( IMF_1 \) containing the finest scale or the shortest period (highest frequency) component of the signal was separated from the rest of the speech signal producing the residue signal \( r_1[k] \):

\[
r_1[k] = x[k] - IMF_1[k]
\]  
(5-28)

Regarding \( r_1[k] \) as a new signal and repeating the sifting process, the second IMF \( IMF_2[k] \) was obtained. Similarly, a series of intrinsic mode functions \( IMF_j[k] \) \((j=1,2,\ldots,n)\) and the final residue \( r_n[k] \) were calculated. The process stopped when \( r_n[k] \) become a monotonic function from which no more IMFs could be extracted.

After performing the IMF decomposition, the initial speech signal could be reconstructed as the following sum of its IMF components and the final residue:

\[
x[k] = \sum_{j=1}^{n} IMF_j[k] + r_n[k]
\]  
(5-29)

For each IMF vector, the Renyi entropy \( ER_j \) of order \( q \) was calculated as:

\[
ER_j = \frac{1}{1-q} \ln \left( \sum_{k=1}^{n} (IMF_j[k])^q \right) \quad k = 1,\ldots,n
\]  
(5-30)
The following averaged ER values formed the feature vectors representing speech frames:

\[ \hat{ER}_j = \frac{1}{n} \sum_{t=1}^{n} (ER_{jt}) \]  

(5-31)

### 5.3.2 Experiments and results for the EMD-AER features

The stress and emotion classification tests were conducted using EMD-AER features. The modeling and classification was performed using two types of classifiers: GMM and KNN.

The speech data for stress classification included datasets from the Speech under Actual Stress domain of the SUSAS database described in Chapter 3 and representing three levels of stress: high, low and neutral. The speech data for emotion classification included dataset from the ORI database described in Chapter 3 and representing five different emotions: neutral, anger, anxious, dysphoric, and happy.

For each feature/classifier combination, the classification process was repeated 15 times, each time with different randomly selected training (80%) and testing (20%) data sets. An average percentage of the identification accuracy (APIA%) was then calculated over 15 repetitions using the formula given in Eq.(3-9).

The results were generated for two different values of the Renyi entropy order: \( q=2 \) and \( q=3 \).

**(A) Results of stress classification using EMD-AER**

Table 5-16 shows APIA% values obtained for stress classification using EMD-AER features and two classifiers GMM and KNN. Figure 5-29 shows the same results but in a form of a bar-graph.

**Table 5-16** APIA% for stress classification using EMD-AER features and the GMM and KNN classifiers.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( q=2 )</th>
<th>( q=3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
</tr>
<tr>
<td>ē vowel</td>
<td>60.61</td>
<td>63.03</td>
</tr>
</tbody>
</table>
CHAPTER 5. NEW FEATURE EXTRACTION METHODS BASED ON SPEECH SPECTROGRAMS, WAVELET PACKET AND EMPIRICAL MODE DECOMPOSITION

<table>
<thead>
<tr>
<th></th>
<th>56.18</th>
<th>53.64</th>
<th>56.73</th>
<th>54.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>ā vowel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>single vowels</td>
<td>53.36</td>
<td>48.79</td>
<td>52.46</td>
<td>50.47</td>
</tr>
<tr>
<td>SUAS</td>
<td>54.69</td>
<td>48.68</td>
<td>54.47</td>
<td>48.08</td>
</tr>
</tbody>
</table>

**Figure 5-29** APIA% for stress classification using EMD-AER features with the Renyi’s entropy of order q=2 and q=3.

Table 5-17 shows an example of the confusion table (correct classification and misclassification rates of individual stress levels) using the GMM classifier and the EMD-AER features (with Renyi entropy q=2).

### Table 5-17

<table>
<thead>
<tr>
<th>Actual level of stress</th>
<th>ē vowel</th>
<th>ā vowel</th>
<th>single vowels</th>
<th>SUAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H</td>
<td>L</td>
<td>N</td>
<td>H</td>
</tr>
<tr>
<td>H</td>
<td>70.77</td>
<td>23.85</td>
<td>5.38</td>
<td>64.29</td>
</tr>
<tr>
<td>L</td>
<td>27.86</td>
<td>62.14</td>
<td>10.00</td>
<td>38.18</td>
</tr>
<tr>
<td>N</td>
<td>15.00</td>
<td>50.00</td>
<td>35.00</td>
<td>38.33</td>
</tr>
</tbody>
</table>

(B) Results of emotion classification using EMD-AER

Table 5-18 shows APIA% values for emotion classification using EMD-AER features and two classifiers GMM and KNN. Figure 5-30 presents the same results as Table 5-18 but in a form of a bar-graph.

### Table 5-18

<table>
<thead>
<tr>
<th></th>
<th>62.00</th>
<th>30.83</th>
<th>7.17</th>
</tr>
</thead>
</table>
CHAPTER 5. NEW FEATURE EXTRACTION METHODS BASED ON SPEECH SPECTROGRAMS,
WAVELET PACKET AND EMPirical MODE DECOMPOSITION

<table>
<thead>
<tr>
<th>Dataset</th>
<th>q=2</th>
<th>q=3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
</tr>
<tr>
<td>ORI</td>
<td>48.00</td>
<td>40.33</td>
</tr>
</tbody>
</table>

**Figure 5-30** APIA% for emotion classification using EMD-AER features with the Renyi’s entropy of order $q=2$ and $q=3$.

Table 5-19 shows an example of the confusion table (correct classification and misclassification rates of individual emotions) using the GMM classifier and the EMD-AER features (with Renyi entropy $q=2$).

**Table 5-19** Confusion table for emotion classification using EMD-AER features (with Renyi entropy $q=2$) and the GMM classifier.

<table>
<thead>
<tr>
<th>Actual emotions</th>
<th>Classified emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Angry</td>
</tr>
<tr>
<td>Angry</td>
<td>31.67</td>
</tr>
<tr>
<td>Anxious</td>
<td>5.00</td>
</tr>
<tr>
<td>Dysphoric</td>
<td>5.00</td>
</tr>
<tr>
<td>Neutral</td>
<td>20.00</td>
</tr>
<tr>
<td>Happy</td>
<td>28.33</td>
</tr>
</tbody>
</table>

**(C) Discussions of results for the EMD-AER features**

The classification results listed in Tables 5-16 and Table 5-18 showed that GMM and KNN classifiers achieved the similar recognition accuracy. The order of the Renyi entropy has little effect on the classification accuracy.

Since the ORI database contains speech recorded in a natural environment during typical conversation, it was expected that the recognition rates should be lower than in the case of the SUSAS data which contains expression produced under highly stressful
conditions (rollercoaster ride, pilot cabin). Although the 40% rate was above the chance level (20%) for 5 emotional classes, more work needs to do to achieve accuracy rates comparable with those provided by the Teager energy operator and the AUSEEG and AUSEES features described in Chapter 6.

Finally, it was observed that the EMD-AER based features showed worse results than the spectral features listed in section 5.1. The low performance of the EMD-AER features can be attributed to the fact that during the estimation of IMFs, the EMD analysis provides low resolution at the high frequency range, which could lead to the loss of important high frequency cues.
Chapter 6. New Features Based on Nonlinear Models of Speech Production

This Chapter introduces a number of new feature extraction methods for an automatic detection of stress and emotion in speech. The proposed feature extraction methods were inspired by a number of recent laryngological experiments and new nonlinear models of speech production. The proposed features were tested using the same speech data sets and classifiers as those used with classical features tested in Chapter 4.

6.1 Introduction: nonlinear model of the glottal flow formation

The classical source-filter model has a linear character and was generated a few decades ago for the purposes of telecommunication engineering, where conveying of an accurate linguistic content was of primarily importance. It does not include mechanisms explicitly responsible for the generation of the paralinguistic aspect of speech. As a result the majority of the current approaches to emotional speech analysis rely on the assumption that the emotional state of a speaker affects in some way speech parameters assumed by the existing source-filter model. Subsequently these parameters including the fundamental frequency $F_0$, formants and energy, or parameters derived from them, are the most often cited in the literature as characteristic features used in emotion recognition from speech [8,9]. An increasing number of recent laryngological and psychological studies aim to improve our understanding of mechanisms involved in speech production and in particular the generation of the paralinguistic aspects of speech [219, 220].

In his original report, Teager [214] presented results of the intra-oral air velocity measurements made with an array of hot wire anemometers. The results indicated that the air flow is not purely laminar and a flow separation occurs causing an active sound generation in the mouth. These findings were later confirmed through extensive experimental studies of fluid flow in a dynamic mechanical model of the vocal folds and tract [221, 222]. The experimental evidence strongly pointed towards the existence of a
vortex train caused by phonation and interacting with the vocal tract boundaries. The proposed theory of speech generation assumed that a speech sound is generated by a combination of sources.

In a study of speech classification under stress Zhou et al. [91] proposed that in the emotional state of anger or stress, additional sound sources can be generated in the form of air vortices. Features sensitive to the presence of these additional vortices can indicate the emotional state of a speaker.

In recent experimental studies [75, 223] using excised canine larynges, two types of consistent, periodic air vortices were identified. During the early opening phase of the vocal folds, when the glottis is convergent, supraglottal vortices occur above the vocal folds. During the latter part of the vocal fold closing, when the glottis is divergent, intraglottal vortices are formed between the vocal folds. The intraglottal vortices generated between symmetrically vibrating vocal folds produced a negative pressure, resulting in a suction force promoting a rapid closing of the vocal folds. The acoustic consequence of the rapid flow shutoff is an increase of energy in the higher harmonics when compared to asymmetrically vibrating vocal folds. The supraglottal vortices on the other hand provide additional sound sources when hitting hard surfaces of the vocal tract or interacting with each other. The presence of these additional sound sources is manifested as additional harmonics and cross-harmonics in the speech spectra. Figure 6-1 shows a new nonlinear model of the air flow during the phonation process based on findings reported in [91] and [75, 223]. Note that the supraglottal vortices do not necessarily coincide in time with the intraglottal vortices.

In Khosla [223], a microphone was placed 8 inches downstream of the vocal folds to record the sound. The results showed a strong correlation between the degree of symmetry of the vocal folds vibration and the distribution of the acoustic energy across frequencies. A relative increase of energy of the high frequency harmonics was observed in the case of non-symmetric vocal folds when compared to the symmetric vocal folds. Assuming that the same holds true for humans, it can be hypothesized that an emotional
state of a speaker can alter the viscosity and elasticity of the vocal folds vibration providing a shift in the spectral energy distribution of the glottal waveform.

![Diagram of glottal flow formation](image)

**Figure 6-1** New nonlinear model of the glottal flow formation based on [91] and [75, 223].

### 6.2 TEO based feature extraction

The non-linear and multi-source approach to speech modelling initiated by Teager opened ways for the extraction of new types of features related to different stressful and emotional states of speakers. It also was noticed by Teager, [214, 224, 225] that energy plays an important function in hearing and recognition of speech.

A number of authors, in their studies of speech classification under stress [91, 226, 227] proposed that in the emotional state of anger or stress, the majority of sound is produced by the laminar flow oscillating with the pitch frequency $F_0$, however the fast air flow also generates air vortices providing additional excitation signals. The additional excitation signals appear in the speech spectrum as harmonics of different fundamental frequencies, not equal to $F_0$, and cross-harmonics between the $F_0$ source and additional sources. The presence of additional harmonics other than the $F_0$-series can therefore indicate the stressful or emotional state of a speaker, and provide a source of characteristic features for the detection and classification of speech under stressful or emotional situations.
In [91], the Teager Energy Operator (TEO) was used to derive a new type of features called the normalised area under envelope of TEO autocorrelation contour (TEO-AutoEnv). The TEO-based feature extraction was performed at the voiced speech level; and the additional harmonics were searched for around F0, and within critical bands[91].

In this study a similar approach was undertaken, in that the Teager energy operator was used to derive characteristic features at the voiced frame level, however the additional harmonics were searched for within critical bands, as well as in wavelet and wavelet packet bands.

### 6.2.1 Teager Energy Operator

In the light of the recent non-linear models of speech production, the speech signal could be regarded as an effect of amplitude and frequency modulation of separate oscillatory waves and modeled as a combination of several amplitude and frequency modulated (AM-FM) oscillatory components. Maragos [215] proposed a nonlinear model of speech, which represents a discrete-time speech signal s[n] as a sum of M components,

\[
s[n] = \sum_{i=1}^{M} x_i[n]
\]  

(6-1)

Each component of speech can be modeled as an AM-FM sine wave given in the discrete time domain as,

\[
x[n] = a[n]\cos(\Phi[n]) = a[n]\cos(\omega_c n + \omega_h \kappa q[k]dk + \theta)
\]  

(6-2)

Where \(a[n]\) is the instantaneous amplitude, \(q[k]\) is the modulating signal, \(\omega_c\) is the source frequency (carrier), \(0 < \omega_h \leq \omega_c\) is the maximum frequency deviation, and \(\theta\) is a constant phase offset. Assuming the above AM-FM modulation of speech, Kaiser [226] proposed the following estimate of the speech instantaneous energy known as the Teager energy operator (TEO):

\[
\Psi[x(t)] = \left[ dx/dt \right] - x(t) \left( d^2 x(t)/dt^2 \right)
\]  

(6-3)
In the discrete-time domain Eq.(6-3) becomes:

\[
\Psi(x[n]) = x[n] - x[n+1]x[n-1] \tag{6-4}
\]

When applying Eq.(6-4) to a discrete time AM-FM signal in Eq.(6-2), the TEO can be expressed as:

\[
\Psi(x[n]) = (a[n])^2 \sin(\omega[n]) \tag{6-5}
\]

Equation Eq.(6-5) indicates that TEO tracks not only the instantaneous amplitude \(a[n]\) of an AM-FM signal but also the instantaneous frequency values \(\omega[n]\).

The great advantage of the TEO analysis is that the classical harmonic analysis based on the spectral energy distribution investigates energy averaged across all samples that belong to a given frame, whereas the TEO energy contour tracks the instantaneous changes in the harmonic structure of a signal.

### 6.2.2 Normalized TEO autocorrelation function

For a speech signal containing only one excitation source with the fundamental frequency \(F_0\), there will be a single harmonic series of integer multiples of \(F_0\). Additional excitation sources other than the vocal folds will generate their own harmonic series.

As suggested in [91], if the speech signal is broken into small bands, and the TEO is calculated for each band, it is easier to observe the presence or absence of additional harmonic component within each band. Moreover, the speech analysis becomes more robust if the characteristic features are derived not directly from the TEO but from the normalized TEO autocorrelation function given as:

\[
R_{\psi_{\psi}}[k] = \frac{1}{2M+1} \sum_{n=-M}^{M} \Psi(x[n])\Psi(x[n+k]) \tag{6-6}
\]

Where \(M\) is the number of samples within the analyzed speech frame.
As illustrated in Figure 6-2, in the simplest case of a single harmonic with a constant instantaneous amplitude and constant instantaneous frequency, the normalized TEO autocorrelation function of Eq.(6-6) for a single frame with N samples will produce a straight line decaying from the point (0,1) to the point (M,0). The area under the autocorrelation envelope in this case, will be equal to M/2. If the analyzed speech frame due to stress or emotion contains more than one harmonic component, the normalized autocorrelation function will produce a time varying envelope decaying in an oscillatory way to zero. The area under the autocorrelation contour in this case will be less than M/2. This indicates that, the area under the normalized TEO autocorrelation contour can be used to detect changes in the harmonic structure of speech caused by conditions such as stress or emotion.

![Figure 6-2](image)

**Figure 6- 2** An example feature using the area under TEO autocorrelation contour.

### 6.2.3 TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-P features

A general flowchart of the feature extraction based on the Teager energy operator is illustrated in Figure 6-3.

![Figure 6-3](image)

**Figure 6- 3** Flowchart of the TEO based feature extraction process.
As showed in Figure 6-3, the area under the normalized TEO autocorrelation contour was calculated on the frame-by-frame basis within relatively narrow frequency bands. Different TEO based approaches to the feature extraction used different ways of the bandwidth subdivision including: critical bands, discrete wavelet transform bands, wavelet packet bands, and perceptual wavelet packet bands. The frequency ranges for these different subdivisions are listed in Table 6-1.

Table 6-1 Frequency bands for different types of auditory scales.

<table>
<thead>
<tr>
<th>No</th>
<th>Critical Bands Filters</th>
<th>Discrete Wavelet Transform</th>
<th>Wavelet Packet</th>
<th>Perceptual Wavelet Packet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower &amp; Upper Bandwidth</td>
<td>Lower &amp; Upper Bandwidth</td>
<td>Lower &amp; Upper Bandwidth</td>
<td>Lower &amp; Upper Bandwidth</td>
</tr>
<tr>
<td>1</td>
<td>100 200 100</td>
<td>0 250 250</td>
<td>0 500 500</td>
<td>0 125 125</td>
</tr>
<tr>
<td>2</td>
<td>200 300 100</td>
<td>250 500 250</td>
<td>100 1000 500</td>
<td>125 250 125</td>
</tr>
<tr>
<td>3</td>
<td>300 400 100</td>
<td>500 1000 500</td>
<td>1500 2000 500</td>
<td>375 500 125</td>
</tr>
<tr>
<td>4</td>
<td>400 510 110</td>
<td>1000 2000 1000</td>
<td>2000 2500 500</td>
<td>625 750 125</td>
</tr>
<tr>
<td>5</td>
<td>510 630 120</td>
<td>2000 4000 2000</td>
<td>2500 3000 500</td>
<td>700 875 125</td>
</tr>
<tr>
<td>6</td>
<td>630 770 140</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
<td>875 1000 125</td>
</tr>
<tr>
<td>7</td>
<td>770 920 150</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
<td>1000 1250 250</td>
</tr>
<tr>
<td>8</td>
<td>920 1080 160</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
<td>1250 1500 250</td>
</tr>
<tr>
<td>9</td>
<td>1080 1270 190</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
<td>1500 1750 250</td>
</tr>
<tr>
<td>10</td>
<td>1270 1480 210</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
<td>1750 2000 250</td>
</tr>
<tr>
<td>11</td>
<td>1480 1720 240</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
<td>2000 2250 250</td>
</tr>
<tr>
<td>12</td>
<td>1720 2080 280</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
<td>2250 2500 250</td>
</tr>
<tr>
<td>13</td>
<td>2080 2320 320</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
<td>2500 3000 500</td>
</tr>
<tr>
<td>14</td>
<td>2320 2700 380</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
<td>3000 3500 500</td>
</tr>
<tr>
<td>15</td>
<td>2700 3150 450</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
</tr>
<tr>
<td>16</td>
<td>3150 3700 550</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
<td>3500 4000 500</td>
</tr>
</tbody>
</table>

Based on different ways of the bandwidth sub-division, the following TEO based features were defined:

*TEO-CB* – the area under the normalized TEO autocorrelation envelope calculated for speech signal within Critical Bands (CB). The frequencies corresponding to CB are listed in Table 6-1. These features were introduced in [91] and showed to provide efficient classification of stress. The Critical Bands were extracted from speech using 16 band-pass filters using frequency bands listed in Table 6-1. The ways in which emotions are coded in speech are usually optimized to suit the characteristics of the human auditory perception. For that reason it appears to be essential that the characteristic features reflect the sensitivity of the human auditory system which is usually approximated by the Critical Bands [189]. The width of these bands increases logarithmically with frequency.
**TEO-DWT** – the area under the normalized TEO autocorrelation envelope calculated for speech signal within 5 octave bands of the Discrete Wavelet Transform decomposition (Table 6-1). This scale uses a relatively small number of frequency bands which leads to the spectral resolution rapidly decreasing towards the high frequencies. These features do not take into account human auditory characteristics.

**TEO-WP** – the area under the normalized TEO autocorrelation envelope calculated for speech signal within 8 octave bands of the Wavelet Packet (WP) decomposition (Table 6-1). In the WP analysis the signal was passed through twice the number of filters used in the DWT analysis, providing additional high frequency bands. Like TEO-DWT, the TEO-WP features do not take into account human auditory characteristics.

**TEO-PWP-S** – the area under the normalized TEO autocorrelation envelope calculated for speech signal within 17 octave bands of the Perceptual Wavelet Packet (PWP) decomposition closely approximating the Critical Bands (Table 6-1). Selections of bands approximating the Critical Bands helped to reflect the human auditory characteristics. Compare to the CB, the PWP decomposition provided frequency bands with improved frequency resolution at the high frequency range.

**TEO-PWP-G** – the area under the normalized TEO autocorrelation envelope calculated for the glottal time waveform within 17 octave bands of the PWP decomposition (Table 6-1). These features use the same bands as TEO-PWP-S, however the decomposition is applied to the glottal waveform rather than speech.

In all features using the wavelet decomposition, mother wavelet db2 was applied. After an extensive comparison of different types of mother wavelets (haar (db1), db2, db3, db5, bior2.4, bior3.1, bior6.8 and coif1), it was found that the db2 mother wavelet provides the best classification performance. The frequency responses of the db2 filters used in the PWP analysis illustrated in Figure 6-5 are showing flat pass-band characteristics.
6.2.4 Experiments and results for the TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G features

Stress and emotion classification tests were conducted using five types of TEO based features: TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G. The modeling and classification was performed using two types of classifiers: GMM and KNN.

The speech data for stress classification included datasets from the Speech under Actual Stress domain of the SUSAS database described in Chapter 3 and representing three levels of stress: high, low and neutral. The speech data for emotion classification included dataset from the ORI database described in Chapter 3 and representing five different emotions: neutral, anger, anxious, dysphoric, and happy.

For each feature/classifier combination, the classification process was repeated 15 times, each time with different randomly selected training (80%) and testing (20%) data sets. An average percentage of the identification accuracy (APIA%) was then calculated over 15 repetitions using the formula given in Eq.(3-9).

(A) Results of actual stress data classification using TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G features
Table 6-2 shows the average percentage of correct classification accuracy (APIA\%) using all five TEO based features and two classifiers GMM and KNN. Figures 6-5 and 6-6 show the APIA\% in a form of bar graph for the GMM and KNN classifier respectively.

### Table 6-2
APIA\% for stress classification using TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G features combined with the GMM and KNN classifiers.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TEO-CB</th>
<th></th>
<th></th>
<th>TEO-DWT</th>
<th></th>
<th></th>
<th>TEO-WP</th>
<th></th>
<th></th>
<th>TEO-PWP-S</th>
<th></th>
<th></th>
<th>TEO-PWP-G</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td></td>
</tr>
<tr>
<td>ē</td>
<td>91.52</td>
<td>86.36</td>
<td>77.58</td>
<td>76.97</td>
<td>77.58</td>
<td>76.67</td>
<td>86.97</td>
<td>86.36</td>
<td>90.61</td>
<td>87.58</td>
<td>90.61</td>
<td>87.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ā</td>
<td>84.18</td>
<td>84.91</td>
<td>72.00</td>
<td>69.27</td>
<td>89.45</td>
<td>90.36</td>
<td>90.36</td>
<td>92.73</td>
<td>86.91</td>
<td>87.82</td>
<td>86.91</td>
<td>87.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>87.20</td>
<td>90.13</td>
<td>74.91</td>
<td>80.39</td>
<td>88.41</td>
<td>91.25</td>
<td>92.76</td>
<td>91.90</td>
<td>82.76</td>
<td>84.40</td>
<td>92.76</td>
<td>84.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUAS</td>
<td>90.94</td>
<td>91.54</td>
<td>77.48</td>
<td>79.03</td>
<td>87.99</td>
<td>90.41</td>
<td>90.94</td>
<td>92.74</td>
<td>86.16</td>
<td>90.09</td>
<td>86.16</td>
<td>90.09</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6-5** APIA\% for stress classification using TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G features combined with the GMM classifier.

**Figure 6-6** APIA\% for stress classification using TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G features combined with the KNN classifier.
As illustrated in Table 6-2, both classifiers showed very close performance with the GMM providing slightly higher classification rates. It can be also noticed that the TEO-PWP-S features outperformed all other features for all data sets except vowel ē, where they showed only slightly lower performance compare to the TEO-CB features.

As the best performing feature, TEO-PWP-S was used in combination with the GMM classifier to generate an example of a confusion matrix showed in Table 6-3.

Table 6-3 An example of a confusion table for stress classification using TEO-PWP-S feature combined with the GMM classifier; H-high level stress, L-low level stress, N-neutral.

<table>
<thead>
<tr>
<th>Actual stress</th>
<th>Classified stress</th>
<th>ē vowel</th>
<th>ā vowel</th>
<th>mixed single vowels</th>
<th>SUAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H</td>
<td>90.77</td>
<td>5.38</td>
<td>3.85</td>
<td>89.52</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>7.14</td>
<td>88.57</td>
<td>4.29</td>
<td>3.18</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>10.00</td>
<td>15.00</td>
<td>75.00</td>
<td>10.00</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>3.85</td>
<td>94.55</td>
<td>2.27</td>
<td>6.35</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>5.38</td>
<td>93.55</td>
<td>2.80</td>
<td>5.58</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>75.00</td>
<td>88.08</td>
<td>1.86</td>
<td>86.43</td>
</tr>
<tr>
<td>N</td>
<td>H</td>
<td>3.85</td>
<td>94.55</td>
<td>2.27</td>
<td>6.35</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>5.38</td>
<td>93.55</td>
<td>2.80</td>
<td>5.58</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>75.00</td>
<td>88.08</td>
<td>1.86</td>
<td>86.43</td>
</tr>
</tbody>
</table>

(B) Results of actual emotion data classification using TEO-CB, TEO-DWT, TEO-WP and TEO-PWP features

Tables 6-4 shows the emotion classification results using TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G features combined with the GMM and KNN classifiers. Figure 6-7 shows the same results in the form of a bar graph. It can be observed in Table 6-4 that in the case of emotion classification, the best performance was provided by the TEO-PWP-S and TEO-CB features. The differences between the two classifiers GMM and KNN were insignificantly small.

Table 6-4 APIA% for emotion classification using TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G features and the GMM and KNN classifiers.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TEO-CB</th>
<th>TEO-DWT</th>
<th>TEO-WP</th>
<th>TEO-PWP-S</th>
<th>TEO-PWP-G</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>84.67</td>
<td>62.50</td>
<td>66.83</td>
<td>84.67</td>
<td>78.10</td>
</tr>
<tr>
<td>KNN</td>
<td>85.00</td>
<td>63.33</td>
<td>85.83</td>
<td>85.10</td>
<td>79.75</td>
</tr>
</tbody>
</table>
Figure 6-7 APIA% for emotion classification using TEO-CB, TEO-DWT, TEO-WP, TEO-PWP features and the GMM and KNN classifiers.

Table 6-5 shows as an example of a confusion table (percentage of correct classification and misclassification within each class) for the best performing features TEO-PWP-S combined with GMM classifier.

Table 6-5 Confusion table for the ORI database using TEO-PWP-S features and the GMM classifier.

<table>
<thead>
<tr>
<th>CE</th>
<th>TEO-PWP-S</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td></td>
<td>Anger</td>
<td>Anxious</td>
<td>Dysphoric</td>
<td>Neutral</td>
</tr>
<tr>
<td>Anger</td>
<td></td>
<td>84.17</td>
<td>1.67</td>
<td>5.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Anxious</td>
<td></td>
<td>3.33</td>
<td>82.50</td>
<td>5.83</td>
<td>4.17</td>
</tr>
<tr>
<td>Dysphoric</td>
<td></td>
<td>1.67</td>
<td>5.83</td>
<td>87.5</td>
<td>2.50</td>
</tr>
<tr>
<td>Neutral</td>
<td></td>
<td>4.17</td>
<td>5.00</td>
<td>4.17</td>
<td>80.00</td>
</tr>
<tr>
<td>Happy</td>
<td></td>
<td>6.67</td>
<td>1.67</td>
<td>0.00</td>
<td>2.50</td>
</tr>
</tbody>
</table>

(C) Discussion of results for the TEO-CB, TEO-DWT, TEO-WP and TEO-PWP features.

One previously known method [91] TEO-CB, and four newly proposed feature extraction methods: TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G were tested in an automatic stress and emotion classification.

The correct stress classification rates ranged from 69.27% to 92.76% depending primarily on the type of features. The best overall performance for stress classification was achieved while using the TEO-PWP-S features providing APIA% of about 92%. The second best performance was given by the TEO-CB with APIA% of about 90%.
In the case of emotion classification, the correct classification rates ranged from 63.33% to 85.83% also depending primarily on the type of features. The same as in the case of stress classification, the best overall performance for emotion classification was achieved while using the TEO-PWP-S and the TEO-CB features, both providing APIA% of about 85%.

In general the average correct classification rates obtained for the SUSAS data (stress) were higher than the rates obtained for the ORI data (emotions). This is consistent with our observations for all previously tested features in Chapter 5 and as previously explained in Section 5.1.8, can be attributed to the high intensity levels of stress provided by the SUSAS data (recorded during rollercoaster ride or in a pilot’s cabin) and relatively low intensity levels provided by the ORI data (recorded during natural conversation between family members).

Importantly, the features based on the nonlinear model of speech production features tested in this section (TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G) clearly outperformed all of the features based on the linear model of speech production tested in Chapters 4 and 5.

Also, the examples of confusion tables for the TEO-PWP-S show very low levels of misclassification between different stress levels and between different emotions when compared with the confusion levels provided by the classical features based on the linear model of speech production.

### 6.2.5 Experiments and results for Stress and emotion classification using multilayer perceptron and probabilistic neural networks combined with TEO-PWP-S features

TEO-PWP-S feature achieved the best performance for both stress and emotion classification studies. In order to test the effect of the classifiers, in this section, stress and emotion classification tests were conducted using TEO-PWP-S features and two types of classifiers: the probabilistic neural network (PNN) and the multilayer perceptron neural network (MLPNN).
In the experiments described here, the MLPNNs had one hidden layer and three output nodes corresponding to three classes of speech: high-level stressed, low-level stressed and neutral. The number of input nodes varied depending on the type of features used in the training process. The number of input nodes was 16 for the TEO-CB features, 5 for the TEO-DWT features, 8 for the TEO-WP features, and 17 for the TEO-PWP-S features. A standard supervised training procedure was applied to derive an optimal set of hidden layer weights by an iterative minimization of the network response error. The optimal set of weights was then used in the classification process.

Experimental results were obtained using three different optimization procedures: the Fletcher-Powell conjugate gradient (cgf), the scaled conjugate gradient (scg) and the resilient back-propagation (rp).

The speech data for stress classification included datasets from the Speech under Actual Stress domain of the SUSAS database described in Chapter 3 and representing three levels of stress: high, low and neutral. The speech data for emotion classification included dataset from the ORI database described in Chapter 3 and representing five different emotions: neutral, anger, anxious, dysphoric, and happy.

For each feature/classifier combination, the classification process was repeated 15 times, each time with different randomly selected training (80%) and testing (20%) data sets. An average percentage of the identification accuracy (APIA%) was then calculated over 15 repetitions using the formula given in Eq.(3-9).

(A) Results of stress classification

Table 6-6 shows APIA% values obtained for stress classification using TEO-PWP-S feature and two types of neural work classifiers: PNN and MLPNN. Figure 6-8 shows the same results but in a form of bar graph.
CHAPTER 6. NEW FEATURES BASED ON NONLINEAR MODEL OF SPEECH PRODUCTION

Table 6-6 APIA% of stress classification for TEO-PWP-S features using MLPNN and PNN classifiers.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training algorithm for MLPNN</th>
<th>PNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>scg</td>
<td>cgf</td>
</tr>
<tr>
<td>ē vowel</td>
<td>80.57</td>
<td>70.29</td>
</tr>
<tr>
<td>ā vowel</td>
<td>80.67</td>
<td>74.06</td>
</tr>
<tr>
<td>Single vowels</td>
<td>88.76</td>
<td>80.89</td>
</tr>
<tr>
<td>SUAS</td>
<td>85.18</td>
<td>75.43</td>
</tr>
</tbody>
</table>

Figure 6-8 APIA% of stress classification for TEO-PWP-S features combined with MLPNN and PNN classifiers.

(B) Results of emotion classification

Table 6-7 shows the APIA% values obtained for emotion classification using TEO-PWP-S features combined with the PNN and MLPNN classifiers.

Table 6-7 APIA% of emotion classification for TEO-PWP-S features using MLPNN and PNN classifiers.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training algorithm for MLPNN</th>
<th>PNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>scg</td>
<td>cgf</td>
</tr>
<tr>
<td>ORI</td>
<td>74.13</td>
<td>53.70</td>
</tr>
</tbody>
</table>

(C) Discussions

The results indicate that in the case of stress classification both classifiers (PNN and MLPNN) show very similar performance, whereas in the case of emotion classification the PNN classifier shows slightly better performance that the MLPNN.
When comparing the performance of PNN and MLPNN in Tables 6-6&6-7 with the performance of GMM and KNN in Tables 6-2&6-4, it can be observed that both neural network classifiers (MLPNN and PNN) provide worse performance than the GMM and KNN classifiers. A possible reason might be because the performance of neural networks was affected by numerous factors, such as the number of hidden layers, or the choice of training algorithms. It was difficult to set up the most optimal parameters.

The MLPNN performed the best when using the scg or the rp algorithms. The MLPP in combination with the cgf algorithm provided significantly worse results.

### 6.3 Features based on the TEO contour

#### 6.3.1 TEO-Contour

The process of calculating the TEO values followed a procedure described in Section 6.2 [91, 224, 225]. Figure 6-9 illustrates a general flowchart of the processing steps.

![Diagram of TEO-Contour](image)

**Figure 6-9** Calculation of the Teager Energy Operator contour values.

The feature extraction process was performed on the frame by frame basis. The frame length was 256 samples with 50% overlap between frames. The Teager energy operator values were calculated within 17 frequency bands determined by the perceptual wavelet packet analysis (PWP) with db2 mother wavelet. The bands covered the entire spectrum from 0 Hz to 4 kHz and were selected from the entire set of the wavelet packet analysis bands as those which closely approximate the critical bands characterizing human auditory perception. The frequencies of the selected PWP bands were: 0-125Hz, 125-250Hz, 250-375Hz, 375-500Hz, 500-625Hz, 625-750Hz, 750-875Hz, 875-1kHz, 1-
1.25kHz, 1.25-1.5kHz, 1.5-1.75kHz, 1.75-2kHz, 2-2.25kHz, 2.25-2.5kHz, 2.5-3kHz, 3-3.5kHz, and 3.5-4kHz. In the discrete time domain the TEO values were calculated using Eq.(6-4).

In order to reduce the data dimensionality, TEO contours were approximated using cubic splines evaluated at a constant number of 10 points/frame/sub-band. The spline contours were then used to derive a number of features as described in the following section.

### 6.3.2 Definitions of features based on the TEO-Contour

(A) TEO-Contour-Mean

The mean value of the TEO-Contour was calculated for each sub-band within a given speech frame. It represents an average energy value of the signal.

(B) TEO-Contour-Median

The median value of the TEO-Contour was calculated for each sub-band within a given speech frame. The use of the median could eliminate effects of occasional dramatically high or low energy levels that are not representative of the signal in general.

(C) TEO-Contour-Var

The variance of the TEO-Contour was calculated for each sub-band within a given speech frame. The variance indicates the spread around the average value of the signal energy.

(D) TEO-Contour-Min/Max

The minimum/maximum of the TEO-Contour was calculated for each sub-band within a given speech frame. It shows changes in the range of the signal energy values.

(E) TEO-Contour-Diff
CHAPTER 6. NEW FEATURES BASED ON NONLINEAR MODEL OF SPEECH PRODUCTION

The first derivative of TEO-Contour was calculated for each sub-band within a given speech frame, which indicates the rate of changes at each TEO-Contour points.

(F) TEO-Contour-Diff2

The second derivative of TEO-Contour was calculated for each sub-band within a given speech frame, which is the rate of changes of the first derivative of TEO-Contour.

6.3.3 Experiments and results for the TEO-Contour-Mean, TEO-Contour-Median, TEO-Contour-Var, TEO-Contour-Min/Max, TEO-Contour-Diff and TEO-Contour-Diff2

(A) Feature’s sets used in stress and emotion classification tests

Experiments presented in the previous section clearly indicated that characteristic features derived from the Teager energy operator (TEO) have the potential to provide very efficient stress and emotion classification in speech.

In this section, four feature sets (set A, set B, set C and set D) consisting of different combinations of the newly defined features based on the TEO contour, were tested in the automatic stress and emotion classification experiments. The composition of each feature set is presented in Table 6-8.

Table 6-8 The feature combination sets based on the parameters derived from the TEO contour.

| Feature Set A: | TEO-PWP-S |
| Feature Set B: | TEO-Contour-Mean |
| Feature Set C: | TEO-Contour-Mean, TEO-Contour-Median, TEO-Contour-Var, TEO-Contour-Min, TEO-Contour-Max |
| Feature Set D: | TEO-Contour-Mean, TEO-Contour-Median, TEO-Contour-Var, TEO-Contour-Min, TEO-Contour-Max, TEO-Contour-Diff, TEO-Contour-Diff2 |
(B) Results of stress classification for the TEO-contour features

Table 6-9 shows the APIA% values obtained for stress classification using the four feature sets (A, B, C and D) and two different classifiers GMM and KNN. Figure 6-10 and Figure 6-11 show the same results but using bar graphs.

Table 6-9 APIA% for stress classification using TEO-contour based feature sets combined with the GMM and KNN classifiers.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Feature Set A</th>
<th>Feature Set B</th>
<th>Feature Set C</th>
<th>Feature Set D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
</tr>
<tr>
<td>ē vowel</td>
<td>86.97</td>
<td>86.36</td>
<td>60.61</td>
<td>63.84</td>
</tr>
<tr>
<td>ā vowel</td>
<td>90.36</td>
<td>92.73</td>
<td>63.27</td>
<td>62.06</td>
</tr>
<tr>
<td>Single vowels</td>
<td>92.76</td>
<td>91.90</td>
<td>60.34</td>
<td>63.59</td>
</tr>
<tr>
<td>SUAS</td>
<td>90.94</td>
<td>92.74</td>
<td>57.55</td>
<td>65.47</td>
</tr>
</tbody>
</table>

Figure 6-10 APIA% for stress classification using TEO-contour based feature sets combined with the GMM classifier.

Figure 6-11 APIA% for stress classification using TEO-contour based feature sets combined with the KNN classifier.
CHAPTER 6. NEW FEATURES BASED ON NONLINEAR MODEL OF SPEECH PRODUCTION

(C) Results of emotion classification for the TEO-contour features

Table 6-10 shows APIA% values obtained for emotion classification using four feature sets (A, B, C and D) and two different classifiers GMM and KNN. Figure 6-12 shows the same results but using a bar graph.

Table 6-10 APIA% for emotion classification using TEO-contour based feature sets combined with the GMM and KNN classifiers.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Feature Set A</th>
<th>Feature Set B</th>
<th>Feature Set C</th>
<th>Feature Set D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
</tr>
<tr>
<td>ORI</td>
<td>84.67</td>
<td>85.83</td>
<td>75.27</td>
<td>70.15</td>
</tr>
</tbody>
</table>

Figure 6-12 APIA% for emotion classification using TEO-contour based feature sets combined with the GMM and KNN classifiers.

(D) Discussion of results for the TEO-contour features

Tables 6-9 and 6-10 indicate that the TEO-PWP-S features (Feature Set A) clearly outperformed all other sets of features (B, C, and D) in both cases of stress and emotion classification providing up to 92% for stress and 85% for emotions. These results demonstrated that the calculation of the area under the normalized autocorrelation envelope (used in TEO-PWP-S) provides much more stable performance compare to the parameters derived directly from the TEO contour. Under stress or emotions, the TEO profile may contain rapidly occurring multiple harmonics and cross-harmonic terms causing very fast changes in the TEO contour. The area under the normalized TEO autocorrelation can suppress some of these changes while still maintaining general fluctuations.
Within the TEO-contour based features, the TEO-contour-mean method (Set B) achieved the highest classification results (63% for stress and 75% for emotions). Feature set C provided the worst performance.

6.4 Features derived from the spectral energy of glottal waveform (AUSEEG) and the speech (AUSEES)

6.4.1 Definition of the AUSSEG and AUSSES features

In Figures 6-13 and 6-14 examples of the spectral energies of the glottal waveforms are presented. Figure 6-13 shows the examples of glottal spectra for the word “break” from the SUAS data pronounced by the same speaker under different levels of stress. Figure 6-14 shows examples of the glottal spectra for the sentences from the ORI data recorded under different emotions. It should be noted that the examples in Figure 6-14 were calculated using different utterances for each emotion. Since the ORI data represented natural speech it was not practical to search for identical utterances pronounced with different emotions. Despite of this limitation, both figures provide a number of important observations.

Firstly, Figures 6-13&6-14 show that different stress levels and different emotions are characterized by amplitude gradients and distributions of energy across frequency. Secondly, the spectral energy decreases with frequency, however the rate of this decrease differs across emotions and across stress levels. And thirdly, the numbers and values of spectral peaks representing different harmonics also vary across emotions.

Using the above findings, the slopes of the glottal spectral envelope were tested as a possible feature for emotion classification. For each frame, a least squares line was fit to the first 8 harmonics. The classification results based on the slope of the best fit line were relatively poor. This was largely due to difficulties associated with determining frequencies of the harmonic components.
CHAPTER 6. NEW FEATURES BASED ON NONLINEAR MODEL OF SPEECH PRODUCTION

**Figure 6-13** Examples of the spectral energies of the glottal waveforms for the word “break” under different levels of stress pronounced by the same speaker.

**Figure 6-14** Examples of the spectral energies of the glottal waveforms produced with different emotions.

It was found that the value of the area under the spectral envelope provided much more stable parameters. Based on this, two closely related types of features AUSEEG and
AUSEES were proposed. The calculation steps for the AUSEEG and AUSEES methods are illustrated in Figure 6-15. In both cases, the FFT algorithm was applied either to the glottal waveform (AUSEEG) or to the voiced speech signal (AUSEES), and the $20\log_{10}$ of the amplitude spectrum was calculated for each frame. The log amplitude levels below an arbitrary threshold $\zeta = 0\text{dB}$ were set to zero. The value of $\zeta$ was determined experimentally. The entire spectral range of 4 kHz was sub-divided into frequency sub-bands and for each sub-band the area under the spectral envelope was calculated generating feature parameters representing a given frame.

In summary, the AUSEEG and AUSEES features can be defined as follows:

**AUSEEG** - the area under the glottal wave energy envelope calculated within 16 spectral sub-bands of equal width of 250 Hz on a linear scale ranging from 0 Hz to 4 kHz.

**AUSEES** - the area under the speech energy envelope calculated within 16 spectral sub-bands of equal width of 250 Hz on a linear scale ranging from 0 Hz to 4 kHz.

Due to the linear bandwidth sub-division the AUSEES and AUSEEG features do not reflect characteristics of the human auditory system. The performance of both types of features was tested with the bandwidth sub-division based on critical and perceptual wavelet packets (PWP). It was found that the linear bandwidth subdivision provided the best stress and emotion classification results.
6.4.2 Experiments and results using the AUSEEG and AUSEES features

Stress and emotion classification tests were conducted using two types of features: AUSEEG and AUSEES. The modeling and classification was performed using two types of classifiers: GMM and KNN.

The speech data for stress classification included datasets from the Speech under Actual Stress domain of the SUSAS database described in Chapter 3 and representing three levels of stress: high, low and neutral. The speech data for emotion classification included dataset from the ORI database described in Chapter 3 and representing five different emotions: neutral, anger, anxious, dysphoric, and happy.

For each feature/classifier combination, the classification process was repeated 15 times, each time with different randomly selected training (80%) and testing (20%) data sets. An average percentage of the identification accuracy (APIA%) was then calculated over 15 repetitions using the formula given in Eq.(3-9).

(A) Results of stress classification using AUSEEG and AUSEES features

Table 6-11 shows APIA% values obtained for stress classification using AUSEEG and AUSEES features and two different classifiers GMM and KNN. Figure 6-16 shows the same results but using a bar graph.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AUSEEG</th>
<th>AUSEES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
</tr>
<tr>
<td>ē vowel</td>
<td>86.97</td>
<td>76.06</td>
</tr>
<tr>
<td>ā vowel</td>
<td>86.91</td>
<td>80.55</td>
</tr>
<tr>
<td>single vowels</td>
<td>78.32</td>
<td>81.51</td>
</tr>
<tr>
<td>SUAS</td>
<td>81.95</td>
<td>87.11</td>
</tr>
</tbody>
</table>
Figure 6-16 APIA% for stress classification using AUSEEG and AUSEES features.

Table 6-12 and Table 6-13 show the confusion arrays for stress detection using AUSEEG and AUSEES features combined with GMM classifier.

**Table 6-12** An example of a confusion table for stress classification using AUSEES feature combined with the GMM classifier; H-high level stress, L-low level stress, N-neutral.

<table>
<thead>
<tr>
<th>AS</th>
<th>ē vowel</th>
<th>ā vowel</th>
<th>mixed single vowel</th>
<th>SUAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H</td>
<td>L</td>
<td>N</td>
<td>H</td>
</tr>
<tr>
<td>H</td>
<td>83.59</td>
<td>10.26</td>
<td>6.15</td>
<td>88.10</td>
</tr>
<tr>
<td>L</td>
<td>6.19</td>
<td>88.57</td>
<td>5.24</td>
<td>5.45</td>
</tr>
<tr>
<td>N</td>
<td>8.89</td>
<td>23.33</td>
<td>67.78</td>
<td>10.00</td>
</tr>
</tbody>
</table>

**Table 6-13** An example of a confusion table for stress classification using AUSEEG feature combined with the GMM classifier; H-high level stress, L-low level stress, N-neutral.

<table>
<thead>
<tr>
<th>AS</th>
<th>ē vowel</th>
<th>ā vowel</th>
<th>mixed single vowel</th>
<th>SUAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H</td>
<td>L</td>
<td>N</td>
<td>H</td>
</tr>
<tr>
<td>H</td>
<td>90.00</td>
<td>5.38</td>
<td>4.62</td>
<td>83.81</td>
</tr>
<tr>
<td>L</td>
<td>1.43</td>
<td>97.86</td>
<td>0.71</td>
<td>3.18</td>
</tr>
<tr>
<td>N</td>
<td>11.67</td>
<td>33.33</td>
<td>55.00</td>
<td>7.50</td>
</tr>
</tbody>
</table>

(B) Results of emotion classification using AUSEES and AUSEEG features

Table 6-14 shows APIA% values obtained for emotion classification using AUSEEG and AUSEES features and two different classifiers GMM and KNN. Figure 6-17 shows the same results but using a bar graph.
Table 6-14 APIA% for emotion classification using AUSEEG and AUSEES features.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AUSEEG</th>
<th></th>
<th>AUSEES</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
</tr>
<tr>
<td>ORI database</td>
<td>86.93</td>
<td>84.83</td>
<td>85.50</td>
<td>70.20</td>
</tr>
</tbody>
</table>

Figure 6-17 APIA% for emotion classification using AUSEEG and AUSEES features.

Table 6-15 shows examples of the confusion tables for the emotion classification using AUSEEG and AUSEES features combined with GMM classifier.

Table 6-15 An example of a confusion table for emotion classification using AUSEEG and AUSEES features combined with the GMM classifier

<table>
<thead>
<tr>
<th>CE</th>
<th>AUSEEG</th>
<th>AUSEES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anger</td>
<td>Anxious</td>
</tr>
<tr>
<td>Anger</td>
<td>87.33</td>
<td>2.17</td>
</tr>
<tr>
<td>Anxious</td>
<td>0.17</td>
<td>94.17</td>
</tr>
<tr>
<td>Dysphoric</td>
<td>1.17</td>
<td>0.67</td>
</tr>
<tr>
<td>Neutral</td>
<td>1.17</td>
<td>3.17</td>
</tr>
<tr>
<td>Happy</td>
<td>2.50</td>
<td>7.00</td>
</tr>
</tbody>
</table>

(C) Discussions of results for the AUSEEG and AUSEES features

It was also demonstrated (Table 6-11 and Table 6-14) that in both cases (stress and emotion classification), the AUSEEG features representing spectral energy distribution of the glottal waveform provided better classification rates (76%-86% for stress, 84%-86% for emotions) than the AUSEES features representing spectral energy distribution of the speech signal (73%-85% for stress, 70%-85% for emotions). This observation could indicate that most of the emotional aspect of speech is generated during the glottal wave formation, before the vocal tract filtering process.
In general, the classification rates provided by AUSEES and AUSEEG are relatively high when compared to the features based on the linear model of speech production such as pitch, formants, MFCC, glottal time/frequency domain features tested in Chapter 4. At the same time, the AUSEES and AUSEEG results are very close to the results provided by the TEO based features described in Section 6.2.

Inline with other experiments described in this thesis, both classifiers GMM and KNN provided very close results.

Interestingly, the confusion tables for AUSEES and AUSEEG features also are consistent with confusion tables for the TEO based features showing relatively low levels of misclassification between different stress levels and between different emotions when compared with the confusion levels provided by the classical features based on the linear model of speech production.
Chapter 7. Emotion Recognition in Speech of Parents of Depressed and Non-Depressed Adolescents

This chapter investigates an automatic emotion classification in spontaneous speech within two family environments: a family with parents and adolescents diagnosed as clinically depressed and a family with parents and adolescent who are non-depressed. The emotion classification results are also used to determine if the classification rates differ between speakers who are parents of depressed adolescents and speakers who are parents of non-depressed adolescents. The study also investigated the effect of gender on emotion classification by looking at the classification rates obtained for mothers and fathers. The classification experiments used the best performing feature extraction methods (TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG) determined in Chapter 6 and two classifiers: GMM and KNN. The speech data used in this chapter was extracted from the ORI data base and extended to include not 5 but 7 different emotions: contempt, angry, anxious, dysphoric, pleasant, neutral and happy. The speech bandwidth was extended from 4 kHz to 11 kHz.

7.1 Introduction

Majority of the current research tends to be focused on the linguistic part of speech signals. Studies concerning the affective content analysis of speech are still fragmented and spread over a number of disciplines such as psychology, linguistics, forms of biology and medicine, and digital signal processing as a part of engineering [16]. These efforts need to be integrated in the near future to provide a comprehensive understanding of the affect generation and communication mechanisms.

Automatic emotion modeling and classification methods investigated by computer science and digital signal processing experts provide means of testing and consolidating many of the psychological and linguistic theories.
The difficulties encountered in automatic classification can indicate the levels of reality for many of the theoretical assumptions made in the areas of behavioral science and psychology. Various clinical applications of affect analysis in speech have been reported in diagnosis of conditions such as depression [30, 81-85], autism [86], Asperger syndrome [87], Alzheimer [88], schizophrenia and Parkinson's disease [89]. These works demonstrate the importance of speech as a diagnostic signal providing valid information about speaker’s mental and physiological state.

Unlike previous chapters, this chapter will not only test the performance of the feature extraction methods but it will also investigate the psychological aspects of relations between adolescent depression symptoms and parental behavior in emotion expression, as well as the gender factor in the expression of affections.

This chapter aims to investigate the automatic emotion recognition in spontaneous speech signals within normal and clinical family environments. The validity of such examination was based on a number of studies in the area of child psychology indicating that family relationships and the interaction processes between the family members are critical factors in the development and maintenance of depressive symptoms in adolescents. Sheeber et al. [229, 230] suggested that the quality of family interactions is relevant for understanding the development of depressive symptoms in adolescents. An adverse family environment is associated with depressive symptoms of adolescents. It has been reported [229, 230] that depression is inversely related to the level of support, attachment, and approval adolescents experience in the family environment. Gottman and Kroff [231] argued that depressive parental behaviour, such as whining and withdrawal, are particularly predictive of a deterioration of the parent-child relationship over time.

The speech data used in this study was extracted from the soundtrack of video recordings analysed in psychological studies of relations between family support and adolescent depressive symptoms described in Sheeber et al. [229, 230] and Davis et al.
The speech data contained speech of 170 adult speakers (parents) including 95 female speakers and 75 male speakers. It was recorded in a laboratory conditions during natural parent-child conversations.

The emotion classification results were analyzed in order to determine if there are differences in affect expression between parents of depressed adolescents and parents of non-depressed adolescents. It was also investigated if there are differences in affect expression between mothers and fathers in general.

Based on the results of Chapter 4-6 it was determined that the best performing features included: TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG, and therefore, in this chapter only these four feature extraction methods were tested and compared. For the consistency with the previous chapters, two classifiers: GMM and KNN were used. These classifiers were showed to provide very similar results in Chapters 4-6. It was also demonstrated in Section 6.2.6, that GMM and KNN provided better performance than the neural network based methods MLPNN and PNN.

Since the application of an automatic emotion classification described in this chapter goes outside of the communication engineering towards behavioral psychology, the speech data included a wider bandwidth and an increased number of emotional classes.

The speech data used in this chapter was extracted from the ORI data base, however the number of different emotional classes was extended from 5 classes analyzed in Chapters 4-6 to 7 classes representing: contempt, angry, anxious, dysphoric, pleasant, neutral and happy. Also the speech bandwidth was extended from 4 kHz analyzed in Chapters 4-6 to 11 kHz.

### 7.2 Speech database

The speech data was provided by the Oregon Research Institute (ORI) for the purpose of psychological studies of relations between family support and adolescent depressive
CHAPTER 7. EMOTION RECOGNITION IN SPEECH OF PARENTS OF DEPRESSED AND NON-DEPRESSED ADOLESCENTS

symptoms. Details of the data base collection process can be found in Sheeber et al. [229, 230] and Davis et al. [232].

The recordings were conducted in a quiet laboratory room during natural parent-child conversations. Family members were seated a few feet apart and Lapel wireless microphones (model: Audio Technica ATW-831-w-a300) were placed on the participants shirts at the chest level in away that was not restriction their speech.

Audio files of parents of the clinically depressed adolescents, and parents of the non-depressed adolescents were selected for the purpose of this study. The two groups of parents were well matched on many demographic variables such as healthy participants were matched to depressed participants on adolescent age, gender, ethnicity, and the socioeconomic characteristics of their schools.

The recordings were coded by trained observers using the Living-In-Family-Environments (LIFE) coding system [153]. Section 3.1.1 contains detailed description of the discrimination rules used in emotional labelling of speech samples.

Table 7-1 Numbers of speech recordings from the ORI data base. The average length of a recording was 1.5 seconds; PD-parents of depressed adolescents, PND- parents of non-depressed adolescents.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Number of recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>contempt</td>
</tr>
<tr>
<td>Mothers (M)</td>
<td>111</td>
</tr>
<tr>
<td>Fathers (F)</td>
<td>92</td>
</tr>
<tr>
<td>PD</td>
<td>142</td>
</tr>
<tr>
<td>PND</td>
<td>61</td>
</tr>
<tr>
<td>Mothers+Fathers</td>
<td>203</td>
</tr>
</tbody>
</table>

Table 7-1 provides numbers of speech samples for the participants and individual emotions. It shows the numbers of recordings for mothers of both depressed and non-depressed adolescents, fathers of both depressed and non-depressed adolescents, parent of depressed adolescents (PD), parent of non-depressed adolescents (PND), and all mothers and fathers together.
Speech recordings were down-sampled from the original sampling frequency of 44 KHz to 22 KHz providing a speech bandwidth of 11 kHz.

### 7.3 Emotion classification system

A general flowchart of the automatic emotion classification process used in this study is illustrated in Figure 7-1.

![Figure 7-1 General flowchart of the emotion classification system.](image)

The pre-processing stage was applied to the speech samples as described in Section 3.2.1. For the comparison purpose, the feature extraction stage employed four different methods described in Chapter 6: TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG. The classification and modeling was performed using two different classifiers: GMM and KNN.

As described in Chapter 6, the four feature extraction methods: TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG derived from the non-linear model of speech production provided superior performance when compared to all other methods tested in Chapters 4-6. Since the speech bandwidth tested in this Chapter was extended from 4 kHz to 11 kHz, the frequency sub-division used in all four feature extraction methods included additional high frequency bands. The updated frequency bands are listed in Table 7-2.
CHAPTER 7. EMOTION RECOGNITION IN SPEECH OF PARENTS OF DEPRESSED AND NON-DEPRESSED ADOLESCENTS

Table 7-2 The frequency bands and the corresponding bandwidth (Hz) for the TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG methods using 11 kHz speech bandwidth.

<table>
<thead>
<tr>
<th>No</th>
<th>TEO-PWP-S and TEO-PWP-G</th>
<th>AUSEES and AUSEEG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>343.75</td>
</tr>
<tr>
<td>2</td>
<td>343.75</td>
<td>687.5</td>
</tr>
<tr>
<td>3</td>
<td>687.5</td>
<td>1031.25</td>
</tr>
<tr>
<td>4</td>
<td>1031.25</td>
<td>1375</td>
</tr>
<tr>
<td>5</td>
<td>1375</td>
<td>1718.75</td>
</tr>
<tr>
<td>6</td>
<td>1718.75</td>
<td>2062.5</td>
</tr>
<tr>
<td>7</td>
<td>2062.5</td>
<td>2406.25</td>
</tr>
<tr>
<td>8</td>
<td>2406.25</td>
<td>2750</td>
</tr>
<tr>
<td>9</td>
<td>2750</td>
<td>3437.5</td>
</tr>
<tr>
<td>10</td>
<td>3437.5</td>
<td>4125</td>
</tr>
<tr>
<td>11</td>
<td>4125</td>
<td>4812.5</td>
</tr>
<tr>
<td>12</td>
<td>4812.5</td>
<td>5500</td>
</tr>
<tr>
<td>13</td>
<td>5500</td>
<td>6187.5</td>
</tr>
<tr>
<td>14</td>
<td>6187.5</td>
<td>6875</td>
</tr>
<tr>
<td>15</td>
<td>6875</td>
<td>8250</td>
</tr>
<tr>
<td>16</td>
<td>8250</td>
<td>9625</td>
</tr>
<tr>
<td>17</td>
<td>9625</td>
<td>11000</td>
</tr>
</tbody>
</table>

7.4 Experiments and results

The emotion classification algorithms were run 15 times. At each run the entire data set of speech utterances representing a given group of speakers (parents of depressed adolescents (PD), parents of non-depressed adolescents (PND), mothers, fathers or mothers+fathers) was randomly divided into a training set (80% of recordings) and a testing set (20% of recordings). The classification score for all classifiers was calculated as an average percentage of identification accuracy (APIA%) defined as Eq.(3-9). A comparison was made between the performances of four feature extraction methods based on the nonlinear model of speech production: TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG. In order to check the consistency, two different classifiers GMM and KNN were used.
CHAPTER 7. EMOTION RECOGNITION IN SPEECH OF PARENTS OF DEPRESSED AND NON-DEPRESSED ADOLESCENTS

The first stage of the experiment aimed to observe the differences in correct classification rates for the seven emotions (contempt, angry, anxious, dysphoric, pleasant, neutral and happy) between parents of non-depressed adolescents (PND) and the parents of depressed adolescents (PD). In the second stage, the aim was to observe the classification rates between speech data containing all mothers (mothers of depressed adolescents and non-depressed adolescent together) and speech data containing all fathers (fathers of depressed and non-depressed adolescents together). The third stage tested the seven emotions classification accuracy for the whole database.

Table 7-3 compares the average classification (APIA%) results between four feature extraction methods and two different classifiers for all tested data sets (mothers, fathers, PD, PND and mothers+fathers).

Table 7-3 The APIA% for 7 emotions classification using TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG features and two classifiers: GMM and KNN; PD-parents of depressed adolescents, PND-parents of non-depressed adolescents.

<table>
<thead>
<tr>
<th>Group</th>
<th>TEO-PWP-S</th>
<th>TEO-PWP-G</th>
<th>AUSEES</th>
<th>AUSEEG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM KNN</td>
<td>GMM KNN</td>
<td>GMM KNN</td>
<td>GMM KNN</td>
</tr>
<tr>
<td>Mothers</td>
<td>84.59 84.29</td>
<td>81.16 84.89</td>
<td>81.70 69.80</td>
<td>85.17 79.25</td>
</tr>
<tr>
<td>Fathers</td>
<td>81.56 83.23</td>
<td>80.14 80.07</td>
<td>78.96 67.22</td>
<td>84.79 84.86</td>
</tr>
<tr>
<td>PD</td>
<td>76.80 82.50</td>
<td>81.32 83.82</td>
<td>78.07 62.07</td>
<td>80.20 75.67</td>
</tr>
<tr>
<td>PND</td>
<td>80.86 83.44</td>
<td>80.80 85.13</td>
<td>64.52 68.24</td>
<td>71.97 71.83</td>
</tr>
<tr>
<td>Mothers+Fathers</td>
<td>76.56 82.82</td>
<td>80.36 82.15</td>
<td>76.07 71.32</td>
<td>89.95 84.27</td>
</tr>
</tbody>
</table>
CHAPTER 7. EMOTION RECOGNITION IN SPEECH OF PARENTS OF DEPRESSED AND NON-DEPRESSED ADOLESCENTS

Figure 7-2 Emotion classification accuracy (APIA %) for parents of depressed adolescents (PD) and parents of non-depressed adolescents (PND).

Figure 7-3 Emotion classification accuracy (APIA %) for mothers and fathers.

The classification accuracy for individual emotions using four types of feature extraction and two classifiers are presented in Table 7-4. Figure 7-4 and 7-5 show the same results but in the form of bar graphs for PD/PND (Figure 7-4) and gender based (mothers and fathers) (Figure 7-5) respectively.

Table 7-4 Correct recognition rates for individual emotions using TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG features combined with the GMM classifier; Co-contempt, Ag-angry, Ax-anxious, Dy-dysphoric, Pl-pleasant, Ne-neutral, Ha-happy.

<table>
<thead>
<tr>
<th>Groups</th>
<th>TEO-PWP-S</th>
<th>TEO-PWP-G</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Co</td>
<td>Ag</td>
</tr>
<tr>
<td>Mother</td>
<td>90.91</td>
<td>87.86</td>
</tr>
<tr>
<td>Father</td>
<td>84.44</td>
<td>90.67</td>
</tr>
<tr>
<td>PD</td>
<td>90.00</td>
<td>78.00</td>
</tr>
<tr>
<td>PND</td>
<td>81.67</td>
<td>87.86</td>
</tr>
<tr>
<td>Mothers+Fathers</td>
<td>84.00</td>
<td>81.72</td>
</tr>
</tbody>
</table>
### Chapter 7. Emotion Recognition in Speech of Parents of Depressed and Non-Depressed Adolescents

<table>
<thead>
<tr>
<th>Groups</th>
<th>AUSEES</th>
<th>AUSEEG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Co</td>
<td>Ag</td>
</tr>
<tr>
<td>Mother</td>
<td>93.94</td>
<td>87.62</td>
</tr>
<tr>
<td>Father</td>
<td>87.41</td>
<td>87.11</td>
</tr>
<tr>
<td>PD</td>
<td>95.24</td>
<td>80.00</td>
</tr>
<tr>
<td>PND</td>
<td>76.67</td>
<td>85.24</td>
</tr>
<tr>
<td>Mothers+Fathers</td>
<td>92.33</td>
<td>83.68</td>
</tr>
</tbody>
</table>

**Figure 7-4** Correct recognition rates for each emotion for parents of depressed (PD) and non-depressed adolescents (PND) using TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG features and the GMM classifier; Co-contempt, Ag-angry, Ax-anxious, Dy-dysphoric, Pl-pleasant, Ne-neutral, Ha-happy.

**Figure 7-5** Gender based classification accuracy for each emotion using TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG features and the GMM classifier; Co-contempt, Ag-angry, Ax-anxious, Dy-dysphoric, Pl-pleasant, Ne-neutral, Ha-happy.
7.5 Discussion and conclusions

(A) Comparison between different classifiers (GMM and KNN)

In line with the results presented in previous chapters (Chapter 4-6), the results presented in Table 7-3 show that both classifiers (GMM and KNN) provided consistency in trends and very close values of the classification rates.

(B) Comparison between different feature extraction methods (TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG)

Depending on the type of features and the classifiers, the correct classification rates (APIA%) in Table 7-3 range from 62% to 89%. These results are relatively high and confirm the results in Chapter 6 showing that the features based on the nonlinear model provide higher emotion classification rates than the classical features based on the linear model of speech production tested in Chapter 4.

The best performing features were:

- For mothers (female speakers) – AUSEEG
- For fathers (male speakers) – AUSEEG
- For parents of depressed adolescents (PD) – TEO-PWP-G
- For parents of non-depressed adolescents (PND) - TEO-PWP-G
- For all speakers together (mathers+fathers) – AUSEEG

Interestingly, in all cases the features extracted from the glottal waveform (AUSEEG and TEO-PWP-G) outperformed features extracted from the speech waveform (AUSEES and TEO-PWP-S). These results could potentially indicate that the emotional aspect of speech is primarily generated during the glottal wave formation and the emotional changes in vocal tract and lip radiation have a secondary importance.

(C) Comparison between emotion discrimination rates for parents of depressed (PD) and parents of non-depressed adolescents (PND)
A comparison between the classification results (APIA%) for PD and PND based on Table 7-3 shows that for the best performing features (TEO-PWP-G), PD provide higher classification rates than PND. It indicates that the features extracted from PD provided better discrimination between emotions than features extracted from PND.

When looking at the classification rates obtained across individual emotions in Table 7-4 and Figure 7-4, for the best performing features (TEO-PWP-G), it is evident that PDA express their dysphoric (sadness and depression) pleasant and neutral emotions in a stronger way (higher correct classification rates) than PNDA. Whereas, the contempt, angry, anxious, and happy emotions are expressed by the PDA in a more subdued way than by PNDA.

These results are consistent with our previous studies [233] and psychological studies [229, 230, 232]. It can be therefore speculated, that parents of depressed adolescent produce more exaggerated expressions of affect than parents of non-depressed children. Parents of non-depressed children on the other hand are likely to express their emotions in more subtle way with less dramatic differences between emotions.

Although, the validity of these conclusions needs to be further tested, it can be noted that automatic emotion detection has the potential to provide important information for psychological and behavioural analysis.

(D) Comparison between emotion discrimination rates for mothers and fathers

A comparison between the classification results for mothers and fathers in Table 7-3 and Figure 7-3 shows that, for all features/classifier combinations the classification rates were higher for mothers than for fathers. In other words, strong gender based differences in emotion classification accuracy were observed with generally higher correct classification rates obtained for female than for male speakers. Therefore, it is likely that females in general provide easier to discriminate (more exaggerated) expressions of affect than males.
When comparing the classification rates obtained across individual emotions in Table 7-4 and Figure 7-5, for the best performing features in the gender dependent case (AUSEEG), it is evident that females express their contempt, angry, anxious, pleasant and neutral emotions in a stronger way (higher correct classification rates) than males. Whereas, the dysphoric (sadness and depression) and happy emotions are expressed by the females in a more subdued way than by males.

(E) Effect of extended bandwidth and number of emotional classes

The best results (APIA%=89% for GMM, APIA%=84% for KNN) obtained for the whole data base (mothers+fathers) and 7 emotions in Table 7-3 when using the AUSEEG features and the GMM and KNN classifiers with the 11 kHz bandwidth can be compared with the ORI results obtained with the same data and the same classifiers but using only 5 emotions and 4 kHz bandwidth in Table 6.14 (APIA%=86% for GMM, APIA%=84% for KNN). It can be observed that the increase number of classes did not provide any detrimental effect on the classification results. The slight increase of the classification rates for the GMM classifier when using an increased number of classes could be due to extended bandwidth of speech used in this chapter. The effect of extended bandwidth to the emotion classification was also discussed in section 8.3.

The results presented in this chapter are generally consistent with some comments from previous psychological studies [229, 230, 232], although this is a preliminary study and more work needs to be done to establish clear methods of experimental designs and the ways in which the results should be interpreted.
CHAPTER 8. EXPERIMENTS COMPARING SELECTED LINEAR AND NONLINEAR FEATURE EXTRACTION METHODS

Chapter 8. Experiments Comparing Selected Linear and Nonlinear Feature Extraction Methods

This Chapter compares a number of features based on the nonlinear model of speech production proposed in Chapter 7 with some of the classical features based on the linear model and described in Chapter 6. The nonlinear features include the TEO based features (TEO-CB/DWT/WP/PWP), as well as features representing spectral energy distribution of speech (AUSEES) and glottal waveform (AUSEEG). These features are compared the classical features including: the fundamental frequency $F_0$, formants and MFCC. Two classifiers GMM and KNN were tested for consistency. The experiments used speech under actual stress from the SUSAS database (7 speakers; 3 female and 4 male) and speech with five naturally expressed emotions (neutral, anger, anxious, dysphoric, and happy) from the ORI corpora (71 speakers; 27 female and 44 male). The classification results demonstrated consistency with the nonlinear model of the phonation process indicating that the harmonic structure and the spectral distribution of the glottal energy provide the most important cues for stress and emotion recognition in speech. The frequency range from 2.5 kHz to 4 kHz provided the most discriminative features which were on their own almost as effective in class discrimination as features representing the entire speech bandwidth (4 kHz).

8.1 Introduction

As a biological signal, speech contains a lot of medical diagnostic information and psychological behavioral information, which in comparison with other biological signals, such as for example eeg or eeg, has been very much under-utilized. One of the reasons for this under-utilization of the vital information present in speech is the combination of high complexity and a wide bandwidth of the speech signal, which makes the analysis relatively more complex than in the case of other bio-signals. Another limiting factor is a serious lack of proper modeling and understanding of the speech production process.
The classical source-filter model has a linear character and was generated a few decades ago for the purposes of telecommunication engineering, where conveying of an accurate linguistic content was of primarily importance. It does not include mechanisms explicitly responsible for the generation of the paralinguistic aspect of speech. As a result the majority of the current approaches to emotional speech analysis rely on the assumption that stress or the emotional state of a speaker affects in some way speech parameters assumed by the existing source-filter model. Subsequently these parameters which include the fundamental frequency $F_0$, formants and energy, or parameters derived from them, are the most often cited in the literature as characteristic features used in emotion recognition from speech [16, 235, 236].

An increasing number of recent laryngological and psychological studies aim to improve our understanding of mechanisms involved in speech production and in particular the generation of the paralinguistic aspects of speech [228][236-237].

This thesis follows the results of recent laryngological experiments [75][223][236] investigating the nonlinear characteristics of air flow during the phonation process. Based on the suggestions presented in these reports, new types of characteristic features proposed and applied to stress and emotion recognition in speech were proposed and tested in Chapter 6 and Chapter 7.

This chapter demonstrates that the proposed nonlinear features provide significantly better performance than the conventional emotion recognition method based on $F_0$, formants, the mel frequency cepstral coefficients (MFCC) and the glottal features. Moreover, the results based on the new features indicated that the emotional aspect of speech is likely to be generated mostly during the glottal flow formation, and the spectral distribution of the glottal energy is an important factor for differentiating between emotions or different levels of stress.

It is also demonstrated that the nonlinear features extracted from the frequency range 2.5 kHz to 4 kHz were almost as powerful in stress and emotion discrimination as features extracted from the entire speech bandwidth of 4 kHz.
8.2 Stress and emotion classification experiments

8.2.1 Experimental setup

(A) Speech data

The study used two data bases: the Speech Under Actual Stress (SUAS) dataset from the publicly available SUSAS corpora described in Chapter 3, and a relatively large clinical data collected by psychologists from the Oregon Research Institute (ORI), for emotion recognition also described in Chapter 3.

Since the SUAS data was only available at the sampling rate of 8 kHz (4 kHz bandwidth), the ORI data was also downsampled to the same rate to enable closer comparison.

(B) Feature extraction methods

Stress and emotion classification experiments were conducted for 16 different feature extraction methods, which can be grouped into three categories.

The first category includes features based on the classical linear model of speech production assuming a single sound source and a laminar air flow. Features that belonged to this category included the fundamental frequency $F_0$, formants, MFCC, glottal time domain parameters (GT) and glottal frequency domain parameters (GF).

The second category comprised features sensitive to the presence of additional harmonics present in speech due to the supraglottal vortices assumed by the nonlinear model of the phonation process. It included features representing the area under the normalized TEO autocorrelation envelope calculated within small frequency sub-bands (TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S and TEO-PWP-G). These features were described in details in Section 6.3.

The third category was also designed to detect changes in the harmonic structure and energy indicated by the nonlinear model. However, this type of features was design to
reflect effects of the intraglottal vortices assumed by the nonlinear model of speech production. This category contained the area under the spectral energy envelope calculated for the glottal wave (AUSEEG) and for the speech signal (AUSEES). Depending on the bandwidth subdivision this group included: AUSEEG-LB, AUSEEG-CB, AUSEEG-PWP, AUSEES-LB, AUSEES-CB and AUSEES-PWP. These features were described in details in Section 6.4.

(C) Modeling and classification methods.

Two modeling and classification methods GMM and KNN were used to test the consistency of results.

8.2.2 Tables of results

For each feature/classifier combination, the classification process was repeated 15 times, each time with different randomly selected training (80% of recordings) and testing (20% of recordings) data sets. An average percentage of the identification accuracy (APIA%) was then calculated over 15 repetitions using Eq.(3-9).

The average classification results given as APIA% values are listed in Tables 8-1 through to 8-3. These results were produced using two classifiers: GMM and KNN. Both classifiers provided the same trends and very close values. The t-test (paired two sample for means) showed that the difference between means for GMM and KNN across all feature extraction methods was statistically insignificant (p=0.28, df=15, Pearson correlation=0.96) for stress data from the SUAS corpora, and statistically significant (p=0.04, df=15, Pearson correlation=0.96) for emotions from the ORI data. In the statistically significant case, the GMM classifier provided higher average classification values than the KNN classifier.
CHAPTER 8. EXPERIMENTS COMPARING SELECTED LINEAR AND NONLINEAR FEATURE EXTRACTION METHODS

Table 8-1 APIA% for stress and emotion classification using classical features: F0, formants, MFCC, GT and GF.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>F0</th>
<th>Formants</th>
<th>MFCC</th>
<th>GT</th>
<th>GF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
</tr>
<tr>
<td>Stress</td>
<td>50</td>
<td>44</td>
<td>52</td>
<td>62</td>
<td>59</td>
</tr>
<tr>
<td>Emotion</td>
<td>38</td>
<td>31</td>
<td>52</td>
<td>45</td>
<td>49</td>
</tr>
</tbody>
</table>

Table 8-2 APIA% for stress and emotion classification using TEO based features.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TEO-CB</th>
<th>TEO-DWT</th>
<th>TEO-WP</th>
<th>TEO-PWP-S</th>
<th>TEO-PWP-G</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>KNN</td>
</tr>
<tr>
<td>Stress</td>
<td>91</td>
<td>91</td>
<td>77</td>
<td>79</td>
<td>88</td>
</tr>
<tr>
<td>Emotion</td>
<td>85</td>
<td>85</td>
<td>62</td>
<td>63</td>
<td>67</td>
</tr>
</tbody>
</table>

Table 8-3 APIA% for stress and emotion classification using the AUSEEG and AUSEES features.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AUSEEG-LB</th>
<th>AUSEEG-CB</th>
<th>AUSEEG-PWP</th>
<th>AUSEES-LB</th>
<th>AUSEES-CB</th>
<th>AUSEES-PWP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
<td>KNN</td>
<td>GMM</td>
</tr>
<tr>
<td>Stress</td>
<td>82</td>
<td>87</td>
<td>65</td>
<td>76</td>
<td>75</td>
<td>77</td>
</tr>
<tr>
<td>Emotion</td>
<td>87</td>
<td>85</td>
<td>69</td>
<td>67</td>
<td>75</td>
<td>69</td>
</tr>
</tbody>
</table>

8.2.3 Stress classification rates

The correct stress classification rates (APIA%) are indicated in Tables 8-1 to 8-3. The APIA% values were calculated as an average of a population of 15 observations using Eq.(3-9). The observation values were found to be normally distributed (Shapiro-Wilk test, p>0.05) and the analysis of variance (one-way ANOVA, number of observations =15, confidence interval=95%) was conducted to determine statistical significance between the correct stress classification rates obtained for different feature extraction methods. The results are listed in Table 8-4.

In general, these results indicated that there were statistically significant difference in the stress classification rates provided by the classical features and features based on the nonlinear model of speech production.

The classification results for three levels of stress in Tables 8-1 to 8-3 show that the features based on the parameters introduced by the nonlinear model of speech production clearly outperformed features based on the classical linear model.
The best overall performance was given by the TEO based features with APIA% within 77%-93% range. The TEO features were followed by the AUSEE features with APIA% ranging from 65% to 87%, and the lowest performance was produced by the classical features with APIA% within the 42% to 65% range.

Amongst the TEO features the best performing were the TEO-PWP-S (93% with KNN, 91% with GMM) and TEO-CB (91% with both GMM and KNN).

Within the AUSEE group, the best performing were AUSEEG-LB (87% with KNN and 82% with GMM) and AUSEES-LB (84% with KNN and 77% with GMM).

The best performing classical features were MFCC (65% with KNN and 59% with GMM), formants (62% with KNN and 52% with GMM) and GT (60% with GMM and 51% with KNN).

A comparison between AUSEES and AUSEEG features in Table 8-3 indicated that the AUSEE features calculated for the glottal waveform (AUSEEG-LB, AUSEEG-CB and AUSEEG-PWP) outperformed the AUSEE features calculated for the speech waveform (AUSEES-LB, AUSEES-CB and AUSEES-PWP). The superior performance of
AUSEEG over AUSEES was statistically significant in all cases except AUSEEG-CB against AUSEES-CB.

### 8.2.4 Emotion classification rates

Tables 8-1 to 8-3 show the classification results for five different emotional classes. Each feature extraction method was tested 15 times, the results were found to be normally distributed (Shapiro-Wilk test, p>0.05). The results of statistical analysis based on one-way ANOVA (number of observations =15, confidence interval=95%) are illustrated in Table 8-5.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>-</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>8</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>9</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>10</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>11</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>12</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>13</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>14</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>15</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>16</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>\</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Like in the case of stress classification, the emotion classification showed statistically significant difference in the classification rates provided by the classical features and features based on the nonlinear model of speech production.

The features related to the nonlinear model of speech production (Tables 8-2 to 8-3) again clearly outperformed features based on the classical linear model (Table 8-1).
Unlike in the stress classification, emotion classification did not show statistically significant differences between performances of the TEO based features and the AUSEE features. Both groups produced very close results ranging from 66% to 87% for AUSEE features, and from 62% to 86% for the TEO based features. The classical features again showed significantly worse results.

Within the AUSEE group, the best performing were AUSEEG-LB (87% with GMM and 85% with KNN) and AUSEES-LB (85% with GMM and 70% with KNN).

Amongst the TEO features the best performing were the TEO-PWP-S (86% with KNN, 85% with GMM) and TEO-CB (85% with both GMM and KNN).

The best performing classical features were formants (52% with GMM and 45% with KNN) and MFCC (49% with GMM and 40% with KNN).

Like in the case of stress classification, the emotion classification results in Table 8-3 showed that the AUSEE features calculated for the glottal waveform (AUSEEG-LB, AUSEEG-CB and AUSEEG-PWP) provided higher classification rates than the AUSEE features calculated for the speech waveform (AUSEES-LB, AUSEES-CB and AUSEES-PWP). The higher performance of AUSEEG compared to AUSEES was statistically significant in all cases except AUSEEG-LB against AUSEES-LB, AUSEEG-CB against AUSEES-CB and AUSEEG-CB against AUSEES-PWP (Table 8-5).

Figures 8-1 and 8-2 show the same results as those listed in Tables 8-1 to 8-3, however the graphical representation in the form of bars makes it easier to illustrate the relative differences between different types of feature extraction.
CHAPTER 8. EXPERIMENTS COMPARING SELECTED LINEAR AND NONLINEAR FEATURE EXTRACTION METHODS

8.2.5 Discriminative powers of different frequency bands

It was determined through classification tests that in general the linear equidistant frequency sub-bands provided better classification results than the logarithmically equidistant sub-bands. As previously indicated in [97], characteristic features which provide high resolution at both low and high frequency bands are essential in stress and emotion recognition in speech.

Figure 8-1 APIA % for all features using the GMM classifier.

Figure 8-2 APIA % for all features using the KNN classifier.
In order to determine which frequency bands provide the best discrimination between stress levels and between different types of emotions, energy contributions from different frequency bands across the whole spectrum were first calculated using:

\[ E_{\text{Contribution}} = \sum_{i=1}^{N} E_{\text{Band \_i}} / E_{\text{Total \_i}} \]  

(8-1)

Where, \( E_{\text{Band \_i}} \) denotes the speech energy within a given frequency band for one recording, and \( E_{\text{Total \_i}} \) is the total energy for all bands for one recording. \( N \) denotes the number of utterances in one dataset (For the SUSAS database as described in section 3.1: \( N=1202 \) for high level stressed speech dataset, \( N=1276 \) for moderate level stressed speech dataset, \( N=701 \) for neutral speech dataset. For the ORI database as described in section 3.1: \( N=200 \) for each emotional speech dataset).

The plots of energy contribution \( E_{\text{Contribution}} \) in Figure 8-3 and Figure 8-4 for stress and emotions respectively were then used to select frequency ranges yielding the largest and the smallest diversity between different stress levels and between different emotions. The AUSEEG features calculated only for these selected frequency ranges were then used in combination with the GMM method to classify stress and emotions. The results provided information about discriminative powers of energy based AUSEEG features corresponding to different frequencies (Table 8-6 and Figure 8-5).

Figure 8-3 shows that in the case of different stress levels, the larges diversity between energy contributions occurs within low frequencies ranging from 0 to 250 Hz and at high frequencies ranging from 2.5kHz to 3.5kHz. Similarly, Figure 8-4 shows that in the case of different emotions, the largest diversity between energy contributions from different frequency bands occurs at the low frequency range 0-250Hz and at the high frequency range 2.5kHz-4kHz. The middle range of frequencies from 250Hz to about 2.5kHz does not show any clear differences between stress levels (Figure 8-3) or between emotions (Figure 8-4).
Figure 8-3 An average percentage of the energy contribution by each of the 16 linear bands to the total energy within a frame of the glottal wave for different stress levels. These results were obtained using the AUSEEG features and the GMM classifier.

Figure 8-4 An average percentage of the energy contribution by each of the 16 linear bands to the total energy within a frame of the glottal wave for different emotions. These results were obtained using the AUSEEG features and the GMM classifier.

The classification results for the frequency ranges showing the largest and the smallest diversity between different classes of stress and emotion are listed in Table 8-6 and plotted in Figure 8-5. It can be observed that the frequency range 2.5 kHz to 4 kHz provided the most discriminative features which were on their own almost as effective in class discrimination as features representing the entire bandwidth. In the case of stress classification, the 2.5-4 kHz features yielded classification rates of 79% whereas the whole bandwidth gave 82%. Similarly, in the case of emotion classification, the 2.5 kHz-
4 kHz features yielded classification rates of 79% whereas the whole bandwidth gave 87%. The middle range of frequencies (250 Hz-2.5 kHz) was much less efficient in the classification task resulting in 51% correct classification rates for stress and only 37% for emotions. The combined 0-250Hz and 2.5kHz -4kHz ranges were also found to be quite efficient providing 80% correct classification rates for stress and 83% for emotions.

**Table 8-6** Emotion classification (APIA %) using AUSEEG-LB/GMM within selected band.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>0-0.25kHz</th>
<th>0.25-2.5kHz</th>
<th>2.5-4 kHz</th>
<th>0-0.25kHz+2.5-4 kHz</th>
<th>0-4 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stress</td>
<td>48</td>
<td>51</td>
<td>79</td>
<td>80</td>
<td>82</td>
</tr>
<tr>
<td>Emotion</td>
<td>25</td>
<td>37</td>
<td>79</td>
<td>83</td>
<td>87</td>
</tr>
</tbody>
</table>

**Figure 8-5** Emotion classification (APIA %) using AUSEEG-LB/GMM within selected band.

### 8.2.6 Confusion between classes

Tables 8-7 and 8-8 show examples of confusion matrix between classes for stress and emotions respectively.

Examples of the misclassification results are presented for the MFCC features, as some of the best performing linear features in stress and emotion classification tasks. The TEO-PWP-S features are used as an example of the best performing nonlinear features in the stress classification, and the AUSEEG-LB features are used as the best performing nonlinear features in emotion classification.

In all cases the class modeling was performed using the GMM method.
Generally the examples in Table 8-7 show that the classical MFCC features not only provide relatively low classification rates (52%-64%) across all stress levels compare to the TEO-PWP features, but also yield high misclassification rates. In particular, the MFCC features provide high misclassification between low level stress and neutral speech. These problems clearly disappeared when the nonlinear TEO-PWP-S features were used providing relatively high correct classification rates (86%-93%) for all levels of stress.

A very similar pattern emerged when analysing the emotion misclassification examples in Table 8-8. The MFCC features provided low classification rates (44%-58%) across all emotions, with particularly high confusion between angry and happy, anxious and dysphoric, and dysphoric and neutral emotions. The AUSEEG-LB features eliminated these problems producing high classification rates (81%-94%) for all emotions.

**Table 8-7** Confusion table for different stress levels based on MFCC and TEO-PWP-S features and GMM classifier.

<table>
<thead>
<tr>
<th>Features</th>
<th>Actual stress</th>
<th>Classified stress</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>High stress</td>
</tr>
<tr>
<td>MFCC</td>
<td>High stress</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Low stress</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>15</td>
</tr>
<tr>
<td>TEO-PWP-S</td>
<td>High stress</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>Low stress</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Neutral</td>
<td>2</td>
</tr>
</tbody>
</table>

**Table 8-8** Confusion table for different emotions based on MFCC and AUSEEG-LB features and GMM classifier.

<table>
<thead>
<tr>
<th>Features</th>
<th>Actual emotions</th>
<th>Classified emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>angry</td>
</tr>
<tr>
<td>MFCC</td>
<td>angry</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>anxious</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>dysphoric</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>neutral</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>happy</td>
<td>20</td>
</tr>
<tr>
<td>AUSEEG-LB</td>
<td>angry</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>anxious</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>dysphoric</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>neutral</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>happy</td>
<td>3</td>
</tr>
</tbody>
</table>
8.3 Conclusions

Stress and emotion classification experiments were conducted using 16 different feature extraction methods grouped into three categories: classical methods based on the linear model of speech production, TEO based methods based on the nonlinear model of speech production assumed to be sensitive to the effects of supraglottal vortices, and AUSEE features also based on the nonlinear model but assumed to be sensitive to the effects of intraglottal vortices.

Through the comparison of the classification rates, it was determined that in the case of stress as well as emotion, the nonlinear TEO and AUSEE features clearly outperformed the classical linear features.

In stress classification the best performance was achieved for the TEO based features, which indicates that speech under stress could be predominantly characterised by generation of supraglottal vortices. This is consistent with Zhou et al. [91].

The emotion classification results on the other hand showed that the TEO and AUSEEG features provided very close performance which could indicate the emotional speech is characterised not only by the formation of supraglottal vortices but also mechanisms involved in formation of the intraglottal vortices, which is consistent with Khosla [75] [223].

A comparison between the AUSEE features calculated for the glottal waveform (AUSEEG-LB, AUSEEG-CB and AUSEEG-PWP) and the AUSEE features calculated for the speech waveform (AUSEES-LB, AUSEES-CB and AUSEES-PWP) showed that the glottal features provided higher performance than the speech based features for both stress and emotion. This observation could indicate that the majority of the paralinguistic aspect of speech (stress and emotion) is generated during the glottal wave formation, before the vocal tract filtering process.

Although, the correctness of the above conclusions needs to be confirmed by further theoretical and experimental studies, it can be stated that the results appear to be...
consistent with the nonlinear model of the air flow during the phonation process [75][91][223-225].

This study also lacks investigations into gender based differences in stress and emotion recognition; however this issue was investigated in Chapter 7.

The other shortcoming of this study which includes a limited bandwidth (4 kHz) of the digitized speech was also addressed in Chapter 7. The findings presented in this chapter showed particular importance of the high frequencies (2.5 kHz – 4 kHz) for the emotion classification. A small-scale investigation was conducted to determine if frequencies above 4 kHz would also hold important information for discrimination between different emotions and stress levels. A wide band (11 kHz bandwidth) speech from the ORI database representing 7 emotions: (contempt, angry, anxious, dysphoric, pleasant, neutral and happy) was used in the classification process based on the AUSEEG features and the GMM classifier. The classification results were generated using the AUSEEG features calculated within 5 different frequency bands: 0 Hz to 687 Hz, 687 Hz to 2.75 kHz, 2.75 kHz to 5.5 kHz, 5.5 kHz to 8.937 kHz and 8.937 kHz to 11 kHz. The results presented in Table 8-9 and Figure 8-6 show that the range from 2.75 kHz to about 5.5 kHz provides the most important features for emotion discrimination in speech. When comparing Table 8-6 and Table 8-9, it can be noted that the case of AUSEEG/GMM classification an widening of the analyzed frequency band from 2.5 kHz -4 kHz to 2.7 kHz to 5.5 kHz did not provide any change in the correct classification rates giving in both cases about 79%.

These results lead to an important conclusion indicating that a narrow band speech (4 kHz) can be used to efficiently discriminate between emotions.

**Table 8-9** Emotion classification (APIA %) using AUSEEG/GMM within selected frequency bands.

<table>
<thead>
<tr>
<th>Frequency bands</th>
<th>0-0.687 kHz</th>
<th>0.687-2.75 kHz</th>
<th>2.75-5.5 kHz</th>
<th>5.5-8.937 kHz</th>
<th>8.937-11 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>APIA (%)</td>
<td>34.22</td>
<td>51.45</td>
<td>79.09</td>
<td>24.82</td>
<td>22.70</td>
</tr>
</tbody>
</table>
CHAPTER 8. EXPERIMENTS COMPARING SELECTED LINEAR AND NONLINEAR FEATURE EXTRACTION METHODS

Figure 8-6 Emotion classification using AUSEEG/GMM within selected frequency bands.
Chapter 9. Discussion and Conclusions

This Chapter summarizes the results of this thesis and discusses some of the problems connected with the findings. A performance based ranking list of all stress and emotion classification methods researched in this thesis is given.

9.1 Performance overview of stress and emotion classification methods researched in this thesis


The proposed features were tested against a range of classical features (F_0, formants, MFCC, GT, GF and TEO-CB) previously described in literature. Majority of the previously proposed features with an exception of (TEO-CB) was based on the classical linear model of speech production.

Majority of the feature extraction methods were tested using two classifiers: GMM and KNN. The TEO-PWP-S features were also tested using two neural network classifiers: MLPNN and PNN.

Table 9-1 provides a list of all feature extraction/classification methods tested in this thesis and ranked in the order of their performance (values of the average percentage of the identification accuracy APIA (%)). The blue shade indicates the new methods introduced in the thesis and based on the linear model of speech production, the yellow shade indicates the new methods introduced in the thesis and based on the nonlinear
model of speech production. Feature that are not highlighted were previously used in literature.

Table 9-1 A ranking list of all feature/classifier combinations tested in this thesis using the same sets of data (SUAS for stress and ORI for emotions). Blue shade - new linear methods introduced in the thesis; Yellow shade - new nonlinear methods introduced in the thesis. Feature that are not highlighted were previously used in literature.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Stress Classification: low stress, high stress and neutral (SUAS data with isolated words)</th>
<th>Emotion Classification: neutral, anger, anxious, dysphoric, and happy (ORI data with isolated sentences)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Features Classifier APIA(%)</td>
<td>Features Classifier APIA(%)</td>
</tr>
<tr>
<td>1</td>
<td>TEO-PWP-S KNN 92.74</td>
<td>AUSEEG GMM 86.93</td>
</tr>
<tr>
<td>2</td>
<td>TEO-CB KNN 91.54</td>
<td>TEO-PWP-S KNN 85.83</td>
</tr>
<tr>
<td>3</td>
<td>TEO-CB GMM 90.94</td>
<td>AUSEES GMM 85.5</td>
</tr>
<tr>
<td>4</td>
<td>TEO-PWP-S GMM 90.94</td>
<td>TEO-CB KNN 85</td>
</tr>
<tr>
<td>5</td>
<td>TEO-PWP-S PNN 90.69</td>
<td>AUSEEG KNN 84.83</td>
</tr>
<tr>
<td>6</td>
<td>TEO-WP KNN 90.41</td>
<td>TEO-CB GMM 84.67</td>
</tr>
<tr>
<td>7</td>
<td>TEO-PWP-G KNN 90.09</td>
<td>TEO-PWP-S GMM 84.67</td>
</tr>
<tr>
<td>8</td>
<td>TEO-WP GMM 87.99</td>
<td>TEO-PWP-G KNN 79.75</td>
</tr>
<tr>
<td>9</td>
<td>AUSEEG KNN 87.11</td>
<td>TEO-PWP-S PNN 79.3</td>
</tr>
<tr>
<td>10</td>
<td>TEO-PWP-G GMM 86.16</td>
<td>TEO-PWP-G GMM 78.1</td>
</tr>
<tr>
<td>11</td>
<td>TEO-PWP-S MLPNN (scg) 85.18</td>
<td>TEO-contour features set B GMM 75.27</td>
</tr>
<tr>
<td>12</td>
<td>AUSEES KNN 84.09</td>
<td>TEO-PWP-S MLPNN (scg) 74.13</td>
</tr>
<tr>
<td>13</td>
<td>TEO-PWP-S MLPNN (rp) 83.48</td>
<td>TEO-WP KNN 71.67</td>
</tr>
<tr>
<td>14</td>
<td>AUSEEG GMM 81.95</td>
<td>TEO-PWP-S MLPNN (rp) 70.47</td>
</tr>
<tr>
<td>15</td>
<td>TEO-DWT KNN 79.03</td>
<td>AUSEES KNN 70.2</td>
</tr>
<tr>
<td>16</td>
<td>TEO-DWT GMM 77.48</td>
<td>TEO-contour features set B KNN 70.15</td>
</tr>
<tr>
<td>17</td>
<td>AUSEES GMM 77.26</td>
<td>TEO-contour features set D GMM 67.45</td>
</tr>
<tr>
<td>18</td>
<td>SS-AF-ERB-AE KNN 77.2</td>
<td>TEO-WP GMM 66.83</td>
</tr>
<tr>
<td>19</td>
<td>WP-ALGF-OFS KNN 76.81</td>
<td>TEO-DWT KNN 63.33</td>
</tr>
<tr>
<td>20</td>
<td>WP-ALGF-OFS GMM 76.35</td>
<td>TEO-contour features set C GMM 62.93</td>
</tr>
<tr>
<td>21</td>
<td>TEO-PWP-S MLPNN (cgf) 75.43</td>
<td>TEO-DWT GMM 62.5</td>
</tr>
<tr>
<td>22</td>
<td>SS-ERB-AE KNN 73.43</td>
<td>TEO-contour features set C KNN 59</td>
</tr>
<tr>
<td>23</td>
<td>SS-sigma-pi (ERB, 32ms) KNN 72.96</td>
<td>TEO-contour features set D KNN 58.45</td>
</tr>
<tr>
<td>24</td>
<td>SS-AF-BARK-AE KNN 72.7</td>
<td>TEO-PWP-S MLPNN (cgf) 53.7</td>
</tr>
<tr>
<td>25</td>
<td>SS-sigma-pi (ERB, 48ms) KNN 71.92</td>
<td>SS-ERB-AE GMM 53.4</td>
</tr>
<tr>
<td>26</td>
<td>SS-AF-ERB-AE GMM 70.94</td>
<td>SS-ERB-AE KNN 52.5</td>
</tr>
<tr>
<td>27</td>
<td>SS-ERB-AE GMM 70.63</td>
<td>Formants GMM 51.5</td>
</tr>
<tr>
<td>28</td>
<td>SS-sigma-pi (Bark, 32ms) KNN 70.09</td>
<td>SS-AF-ERB-AE GMM 51.5</td>
</tr>
<tr>
<td>29</td>
<td>SS-BARK-AE KNN 70</td>
<td>SS-BARK-AE GMM 51.4</td>
</tr>
<tr>
<td>30</td>
<td>SS-AF-CB-AE KNN 68.33</td>
<td>SS-BARK-AE KNN 49.83</td>
</tr>
<tr>
<td></td>
<td>Method</td>
<td>Model 1</td>
</tr>
<tr>
<td>---</td>
<td>------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>31</td>
<td>SS-sigma-pi (Bark, 48ms)</td>
<td>KNN</td>
</tr>
<tr>
<td>32</td>
<td>SS-CB-AE</td>
<td>KNN</td>
</tr>
<tr>
<td>33</td>
<td>SS-sigma-pi (ERB, 32ms)</td>
<td>GMM</td>
</tr>
<tr>
<td>34</td>
<td>SS-sigma-pi (ERB, 48ms)</td>
<td>GMM</td>
</tr>
<tr>
<td>35</td>
<td>TEO-contour features set B</td>
<td>KNN</td>
</tr>
<tr>
<td>36</td>
<td>MFCC</td>
<td>KNN</td>
</tr>
<tr>
<td>37</td>
<td>SS-sigma-pi (Bark, 48ms)</td>
<td>GMM</td>
</tr>
<tr>
<td>38</td>
<td>SS-ALGF-OFS</td>
<td>GMM</td>
</tr>
<tr>
<td>39</td>
<td>SS-ALGF-OFS</td>
<td>KNN</td>
</tr>
<tr>
<td>40</td>
<td>SS-BARK-AE</td>
<td>GMM</td>
</tr>
<tr>
<td>41</td>
<td>SS-AF-BARK-AE</td>
<td>GMM</td>
</tr>
<tr>
<td>42</td>
<td>TEO-contour features set D</td>
<td>KNN</td>
</tr>
<tr>
<td>43</td>
<td>SS-sigma-pi (Bark, 32ms)</td>
<td>GMM</td>
</tr>
<tr>
<td>44</td>
<td>SS-sigma-pi (CB, 48ms)</td>
<td>KNN</td>
</tr>
<tr>
<td>45</td>
<td>SS-sigma-pi (CB, 32ms)</td>
<td>KNN</td>
</tr>
<tr>
<td>46</td>
<td>SS-SP-ALGF-OFS</td>
<td>GMM</td>
</tr>
<tr>
<td>47</td>
<td>Formants</td>
<td>KNN</td>
</tr>
<tr>
<td>48</td>
<td>GP-T</td>
<td>GMM</td>
</tr>
<tr>
<td>49</td>
<td>SS-AF-CB-AE</td>
<td>GMM</td>
</tr>
<tr>
<td>50</td>
<td>SS-CB-AE</td>
<td>GMM</td>
</tr>
<tr>
<td>51</td>
<td>MFCC</td>
<td>GMM</td>
</tr>
<tr>
<td>52</td>
<td>TEO-contour features set C</td>
<td>KNN</td>
</tr>
<tr>
<td>53</td>
<td>SS-SP-ALGF-OFS</td>
<td>KNN</td>
</tr>
<tr>
<td>54</td>
<td>TEO-contour features set B</td>
<td>GMM</td>
</tr>
<tr>
<td>55</td>
<td>SS-sigma-pi (CB, 48ms)</td>
<td>GMM</td>
</tr>
<tr>
<td>56</td>
<td>EMD-AER (q=2)</td>
<td>GMM</td>
</tr>
<tr>
<td>57</td>
<td>EMD-AER (q=3)</td>
<td>GMM</td>
</tr>
<tr>
<td>58</td>
<td>SS-sigma-pi (CB, 32ms)</td>
<td>GMM</td>
</tr>
<tr>
<td>59</td>
<td>Formants</td>
<td>GMM</td>
</tr>
<tr>
<td>60</td>
<td>TEO-contour features set D</td>
<td>GMM</td>
</tr>
<tr>
<td>61</td>
<td>TEO-contour features set C</td>
<td>GMM</td>
</tr>
<tr>
<td>62</td>
<td>GP-T</td>
<td>KNN</td>
</tr>
<tr>
<td>63</td>
<td>$F_0$</td>
<td>GMM</td>
</tr>
<tr>
<td>64</td>
<td>EMD-AER (q=2)</td>
<td>KNN</td>
</tr>
<tr>
<td>65</td>
<td>GP-F</td>
<td>GMM</td>
</tr>
<tr>
<td>66</td>
<td>EMD-AER (q=3)</td>
<td>KNN</td>
</tr>
</tbody>
</table>
9.2 Classical versus new nonlinear feature extraction methods

The results presented throughout the thesis and in the Table 9-1 as well provided an overwhelming support for the nonlinear theory of speech production assuming a multi-source character of speech production and occurrence of air vortices contributing to changes in energy and harmonic structure of the glottal wave due to stress and emotions. The feature extraction methods (TEO-CB, TEO-DWT, TEO-WP, TEO-PWP-S, TEO-PWP-G, AUSEES and AUSEEG) derived under the assumption of the nonlinear processes clearly outperformed all feature extraction methods based on the classical linear theory.

9.3 Stress classification

9.3.1 The best performing feature extraction and classification methods in stress classification

In the case of stress classification, the nonlinear features clearly outperformed all of the linear features. In particular, the TEO-PWP-S features were consistently providing very high correct classification rates indicating that the distribution of speech energy across frequency plays important role in distinguishing between different stress levels. The top 5 best performing feature/classifier combinations for stress classification were (Table 9-1):

1. TEO-PWP-S/KNN
2. TEO-CB/KNN
3. TEO-CB/GMM
4. TEO-PWP-S/GMM
5. TEO-PWP-S/PNN
CHAPTER 9. DISCUSSIONS AND CONCLUSIONS

9.3.2 Confusion between different levels of stress

The nonlinear feature extraction methods not only provided higher average values of the correct classification but also improved correct classification rates for individual stress levels and decreased the misclassification between different stress levels.

9.4 Emotion classification

9.4.1 The best performing feature extraction and classification methods in emotion classification

In the case of emotion classification, the nonlinear features again outperformed all of the linear features. Interestingly, this time the AUSEEG features appear at the to 5 best performing methods, which may indicate that in the case of emotions, the distribution of glottal energy across frequencies plays an important role in distinguishing between different emotions. The top 5 best performing feature/classifier combinations for emotion classification were (Table 9-1):

1. AUSEEG/GMM
2. TEO-PWP-S/KNN
3. AUSEES/GMM
4. TEO-CB/KNN
5. AUSEEG/KNN

9.4.2 Confusion between different types of emotions

Like in the case of stress, for emotion classification, improved correct classification rates for individual types of emotions decreasing the misclassification between different stress levels when compared to the linear features.

In general, the examples of the confusion confusion tables for linear features (in Chapters 4 and 5) show that anger was most often confused with happiness and the anxious emotion was most often confused with the dysphoric emotional state. The
nonlinear methods (as described in Chapter 6) show consistently low level of misclassification between all types of emotions.

### 9.5 Type of classifier

Majority of the experiments used the GMM and KNN classifiers. The TEO-PWP-S features were also tested using two neural network based classifiers: MLPNN and PNN.

It was found that the GMM and KNN classifiers provided very close results and outperformed both of the neural network classifiers.

Amongst the neutral network classifiers the PNN provided generally better results than the MLPNN.

### 9.6 Stress versus emotions

The results in Table 9-1 show that generally, higher correct classification rates were obtained for the stress classification when using the SUSAS data base; the emotion classification based on the ORI data provided lower classification rates. Any direct comparison between results obtained for stress and emotion classification have to be taken with caution due to the following reasons:

- Stress and emotions generally represent different psycho physiological phenomena; although the mechanisms linking these processes with the speech production are not known it is highly likely that they could lead to differences in the automatic detection efficiency.

- The stress classification was performed using only 3 classes, whereas the emotion classification was conducted using either 5 or 7 classes.

- The stress classification was based on isolated words, whereas the emotion classification was based on isolated sentences.

- The speech samples representing different levels of stress were produced under highly stressful conditions (rollercoaster and pilot’s cockpit) when the arousal
levels can be expected to be very high leading to easier differentiation between different levels of stress. The emotional speech, on the other hand, was recorded during a typical discussion, when emotions are usually expressed with low or mild levels of emotional arousal making the differentiation between different emotions more difficult.
References

REFERENCES


186
REFERENCES


63. Scherer C., Can automatic speaker verification be improved by training the algorithms on emotional speech. ICSLP 2000, 2000.
REFERENCES


REFErences

REFERENCES


