Contents

1 Introduction 1

1.1 Motivation for current topic of research 4

1.2 Thesis Objective 6

1.3 Thesis Organisation 7

1.4 Thesis Statement 9

1.5 Thesis Contributions 9

2 Synthesis of dialectal speech: A Survey 11

2.1 Synthesis of Speech 12

2.1.1 Formant Synthesis 12

2.1.1.1 Rule-based Formant Synthesis 15

2.1.2 Concatenative Synthesis 16

2.1.3 Articulatory Synthesis 18

2.1.4 Statistical Parametric Synthesis 20

2.1.5 DNN based Speech Synthesis 21

2.2 Bringing naturalness to synthesised speech 22

2.2.1 Prosody Modelling 22

2.2.2 Voice Source (VS) Modelling 23

2.2.3 Voice Conversion (VC) for expressive speech 24

2.2.4 Formant modelling 26

2.3 Synthesis of Dialectal Speech 27

2.4 Cross-Dialectal Differences 29

2.5 Limitations of current approaches towards the Generation of Dialectal Speech 32

3 A Quest for Dialect-Distinctive Features 35

3.1 Introduction . . . 35

3.1.1 Assamese and its dialects . . . 36

3.1.2 The Assamese script . . . 37

3.1.3 Linguistic variations in Assamese dialects . . . 38

3.1.4 A preliminary analysis of two varieties of Assamese . . . 41

3.2 Analysis of VOT . . . 43

3.2.1 Literature Review . . . 44

3.2.2 Experimental Framework . . . 45

3.2.2.1 Building the Speech Corpus . . . 45

3.2.2.2 Measuring VOT . . . 46

3.2.2.3 Methodology . . . 47

3.2.3 Results and Observations . . . 48

3.3 Analysis of Vowel Space and Duration . . . 53

3.3.1 Literature Review . . . 54

3.3.2 Experimental Framework . . . 55

3.3.2.1 Building of Speech Corpus . . . 55

3.3.2.2 Measuring formants and vowel duration . . . 55

3.3.2.3 Methodology . . . 56

3.3.3 Results and Observations . . . 57

3.4 Analysis of Diphthong formant trajectories . . . 59

3.4.1 Literature Review . . . 59

3.4.2 Experimental Framework . . . 60

3.4.2.1 Building of Speech Corpus . . . 60

3.4.2.2 Formant measurement in Diphthongs . . . 60

3.4.2.3 Methodology . . . 61

3.4.3 Results and Observations . . . 62

3.5 Analysis of Spectral Tilt . . . 64

3.5.1 Literature Review . . . 65

3.5.2 Experimental Framework . . . 65

3.5.2.1 Building the Speech Corpus . . . 65

3.5.2.2 Measuring Spectral Tilt 66
3.5.2.3 Methodology 66
3.5.3 Results and Observations 67
3.6 Analysis of Spectral (MCEP) and Prosodic Features (F0) 67
3.6.1 Literature Review 69
3.6.2 Experimental Framework 69
3.6.2.1 Building of Speech Corpus 69
3.6.2.2 Methodology 69
3.6.3 Results and Observations 72
3.7 Discussion 77

4 Naturalness in Synthesised speech: the Voice Conversion way 79
4.1 Introduction 79
4.2 Voice Conversion 80
4.3 Motivation 81
4.4 Literature Review 83
4.4.1 Dialect Synthesis 84
4.4.2 Accent Conversion 84
4.4.3 Voice Conversion 85
4.4.4 Prosody transformation and Dialects 86
4.5 Methodology 87
4.5.1 System A: The standard HTS 87
4.5.1.1 Speech Database for the HTS system 89
4.5.1.2 Building the HTS 90
4.5.2 System B: The VC System 93
4.5.2.1 Selection of features for conversion 93
4.5.2.2 Choosing a model for the mapping function 94
4.5.2.3 The Speech Processing Toolkit (SPTK) 95
4.5.2.4 Speech Database for the VC system 95
4.5.2.5 Transformation of Spectral features using VQ with Mapping Codebooks 97
4.5.2.6 Transformation of Spectral features using GMM . . 97

4.5.2.7 Transformation of Spectral features using ANN . . 99

4.5.2.8 Transformation of Pitch and Duration 100

4.5.3 Systems C and D: PID Manipulation 101

4.6 Results and Evaluation 102

4.6.1 Results of spectral conversion 102

4.6.2 Results of F0 conversion 105

4.6.3 Results of duration conversion 106

4.6.4 Subjective evaluation using MOS 107

4.7 Summary 110

5 Naturalness in Synthesised speech: the Formant Transformation way 113

5.1 Introduction 113

5.2 Literature Review 114

5.3 Motivation 116

5.4 Experimental Framework 116

5.4.1 Building the Speech Corpus 117

5.4.2 Methodology 118

5.4.2.1 Analysis of Vowel Space 118

5.4.2.2 Fitting a GMM to formant data 125

5.4.2.3 Development of the Mapping Function 127

5.5 Results and Evaluation 130

5.5.1 Objective Evaluation using RMSE 130

5.5.2 Comparison with the ANN based method of Formant Transformation 133

5.5.3 Graphical representation of results using SSANOVA 138

5.5.4 Graphical Comparison of formant transformation results obtained using VC and FT 145

5.5.5 Subjective Evaluation 152

5.6 Summary 152

6 Conclusions and Future Research Directions 155

6.1 Research Contributions 155

6.2 Future Directions 157

A Text Data for building the HTS 179

B Label files for building the HTS 185