List of Figures

1.1 A simplified diagram of a Text-to-Speech Synthesiser[1] 2

1.2 Block Diagram of the proposed Model 7

2.1 Basic Formant Synthesiser [2] 14

2.2 (a) A Parallel Formant Synthesiser (b) A Cascade Formant Synthesiser [3] 14

2.3 The Klatt Synthesiser [76] 16

2.4 Overview of General Unit Selection scheme (Source:[164]) 18

2.5 Tube Model for Articulatory Speech Synthesis [4] 19

2.6 Block diagram of a HMM-based speech synthesis system (HTS) (Source:[164]) 21

3.1 Linguistic Affiliation of the Assamese language. [5] 36

3.2 Distribution of Assamese speakers in and around the state of Assam 37

3.3 (a) Vowels in Assamese Phoneme Inventory (b) Consonants in Assamese Phoneme Inventory [6] 39

3.4 Classification of Assamese Stop Consonants 43

3.5 Voice Onset Time (source: [158]) 47

3.6 Measurement of VOT (a) vot(/t/)=10ms (b) vot(/b/)=-129ms . . 48

3.7 Voice Onset Time (min, max, mean) in NAL and AIR 49

3.8 Waveform & Spectrogram for the word ‘baamun’ 50

3.9 Waveform and Spectrogram for the word ‘pujaari’ 51

3.10 Waveform and Spectrogram for the word ‘ghaam’ 52

3.11 Mean VOTs of Nalbaria variety and AIR variety 52

3.12 Vowel Chart and Vowel Space 54

3.13 F2 vs F1 plots for NAL & AIR 57

3.14 Normalised vowel spaces for (a) AIR and (b) NAL 58

3.15 Average Duration of vowels in AIR and NAL 58

3.16 Dynamic F2-F1 plot of a diphthong in (a) AIR & (b) NAL 62

3.17 Diphthong /eu/ in (a) AIR and (b) NAL 62

3.18 Diphthong /aai/ in (a) AIR and (b) NAL 63

3.19 Diphthong /oi/ in (a) AIR and (b) NAL 63

3.20 Distance between nucleus and offglide of some commonly used diphthongs in AIR and NAL 64

3.21 H1, H2, A1, A2 in a vowel segment 66

3.22 Spectral Tilt values in (a) AIR, (b) NAL and (c) TTS speech data . 68

3.23 Pitch contours in (a) AIR and (b) NAL (Declarative) 70

3.24 Pitch contours in (a) AIR and (b) NAL (Interrogative) 71

3.25 Pitch contours in (a) AIR and (b) NAL (Exclamatory) 72

3.26 Mel Cepstral Coefficients in AIR and Nalbaria speech data 73

3.27 Normalised Mel Cepstral Coefficients in AIR and Nalbaria speech data 74

3.28 KLD distance between mixtures in AIR and Nalbaria GMMs 74

3.29 F0(min) for utterances (Declarative, Interrogative, Exclamatory) in AIR and NAL 75

3.30 F0(max) for utterances (Declarative, Interrogative, Exclamatory) in AIR and NAL 75

3.31 F0(mean) for utterances (Declarative, Interrogative, Exclamatory) in AIR and NAL 76

3.32 F0(std.dev.) for utterances (Declarative, Interrogative, Exclamatory) in AIR and NAL 76

4.1 A typical Voice Conversion System. (courtesy:[96]) 82

4.2 Block Diagram Representation of Proposed System 83

4.3 VC for Dialectal Speech? 87

4.4 Block Diagram Representation of Systems A, B, C and D 88

4.5 Phoneset for Assamese 92

4.6 5-layer ANN architecture with 20 input and 20 output nodes 100

4.7 Source vs Target MCEPs of a Test Utterance 103

4.8 Converted vs Target MCEPs of a Test Utterance 103

4.9 Comparison of MCDs with the 3 methods 105

4.10 Source F0 contour vs Target F0 contour 106

4.11 Mapped F0 contour vs Target F0 contour 106

4.12 Transforming vowel/diphthong durations of a Test Utterance 108

4.13 Details of the MOS test 109

5.1 A sample textgrid in PRAAT showing annotations in three tiers vowel, consonant and word. 119

5.2 A screenshot of the Excel sheet showing source formant data at 20%, 40%, 60% and 80% of vowel/diphthong duration. 119

5.3 Mean & Std.Deviation of F1,F2 formant frequencies in Source and Target Vowels/Diphthongs 120

5.4 Normalised F2 vs F1 plot for (a) AIR, (b) NAL, (c) TTS 121

5.5 F2 vs F1 plot (unnormalised) for (a) AIR, (b) NAL, (c) TTS 122

5.6 Distance between vowels in (a) VS-AIR and VS-NAL, (b) VS-TTS and VS-NAL 125

5.7 Diphthongs /ui/ and /eu/ in Source and Target 125

5.8 GMM clusters in Source Vowel Formant Data 126

5.9 GMM clusters in Target Vowel Formant Data 127

5.10 Scatter Plot with GMM contours fitted to (a) Source Vowel Formant Data and (b) Target Vowel Formant Data 128

5.11 Transformation of F1 formant frequencies 131

5.12 Transformation of F2 formant frequencies 132

5.13 Plot showing %improvement in RMSE values after transformation using Approach 1 133

5.14 Plot showing %improvement in RMSE values after transformation using the 3 approaches 133

5.15 F2 vs F1 for Source(s) 134

5.16 F2 vs F1 for Target(t) 134

5.17 F2 vs F1 for Converted(c) 134

5.18 Vowel distance after Formant Transformation using the GMM-based Mapping Function 135

5.19 ANN architecture for formant transformation with 2 hidden layers having 60 neurons each 136

5.20 RMSE values after transformation using (i) ANN-based & (ii) GMM-based methods . . . 137

5.21 RMSE values after transformation using the (i) ANN-based & (ii) GMM-based methods for Utterance 1 . . . 137

5.22 RMSE values after transformation using the (i) ANN-based & (ii) GMM-based methods for Utterance 2 . . . 138

5.23 RMSE values after transformation using the (i) ANN-based & (ii) GMM-based methods for Utterance 3 . . . 138

5.24 Transformation results for a test utterance (utt2) using the ANN-based method . . . 139

5.25 Transformation results for a test utterance (utt2) using the GMM-based method . . . 139

5.26 Smoothing Splines for vowel /ax/ (a) Source vs Target (b) Converted vs Target . . . 140

5.27 Smoothing Splines for vowel /aa/ (a) Source vs Target & (b) Converted vs Target . . . 141

5.28 Smoothing Splines for vowel /i/ (a) Source vs Target & (b) Converted vs Target . . . 141

5.29 Smoothing Splines for vowel /u/ (a) Source vs Target & (b) Converted vs Target . . . 142

5.30 Smoothing Splines for vowel /e/ (a) Source vs Target & (b) Converted vs Target . . . 142

5.31 Smoothing Splines for vowel /o/ (a) Source vs Target & (b) Converted vs Target . . . 143

5.32 Smoothing Splines for diphthong /oi/ (a) Source vs Target & (b) Converted vs Target . . . 143

5.33 Smoothing Splines for diphthong /ou/ (a) Source vs Target & (b) Converted vs Target . . . 144

5.34 Smoothing Splines for diphthong /aai/ (a) Source vs Target & (b) Converted vs Target . . . 144

5.35 Smoothing Splines for diphthong /ui/ (a) Source vs Target & (b) Converted vs Target . . . 145

5.36 Diphthong /iu/ (a) Target, (b) Converted(VC) & (c) Converted(FT) 146

5.37 Vowel /i/ (a) Target, (b) Converted(VC) & (c) Converted(FT) . . . 147

5.38 Vowel /ax/ (a) Target, (b) Converted(VC) & (c) Converted(FT) . . 148

5.39 Diphthong /oi/ (a) Target, (b) Converted(VC) & (c) Converted(FT) 149

5.40 Diphthong /aai/ (a) Target, (b) Converted(VC) & (c) Converted(FT) 150

5.41 Diphthong /axaa/ (a) Target, (b) Converted(VC) & (c) Converted(FT) 151

List of Tables

3.1 List of stop consonants of Assamese 46

3.2 Range and mean of VOTs for NAL−Nalbaria variety and AIR−All India Radio variety 49

3.3 Vowels in AIR and NAL along with their carrier words 56

3.4 Frequently used diphthongs in AIR and NAL along with their carrier words 61

3.5 Range of 'd' values of diphthongs for speakers in AIR and NAL. ('d': Euclidean distance between nucleus and offglide of a diphthong) . . 64

4.1 Comparison of MCD values before and after conversion of MCEPs via the VQ method 104

4.2 Comparison of MCD values before and after conversion of MCEPs via the GMM method 104

4.3 Comparison of MCD values before and after conversion via the ANN method 104

4.4 Comparison of RMSE(F0) values 107

4.5 Mean and Standard Deviation(StdDev) of duration of vowels in Source and Target data 107

4.6 MOS Results 109

List of Acronyms

AAE African American English

AF Amplitude of frication

AH Amplitude of aspiration

AIR All India Radio

ANN Artificial Neural Network

ANOVA Analysis of Variance

ATR Advanced Telecommunications Research

AV Amplitude of voicing

A1 Amplitude of first formant

A2 Amplitude of second formant

BW Bandwidth

BFGS Broyden-Fletcher-Goldfarb-Shanno

CPP Cepstral Peak Prominence

CTD Central Thai Dialect

CLSP Center for Language and Speech Processing

CMU Carnegie Melon University

DCT Discrete Cosine Transform

DFW Dynamic Frequency Warping

DMOS Differential Mean Opinion Score

DNN Deep Neural Network

DRT Diagnostic Rhyme Test

DTW Dynamic Time Warping

EGG Electro Glottograph

EM Expectation Maximization

EMA Electromagnetic Acoustic Imaging

FFT Fast Fourier Transform

F0 Fundamental Frequency (Hz)

F1 First Formant Frequency (Hz)

F2 Second Formant Frequency (Hz)

F3 Third Formant Frequency (Hz)

FT Formant Transformation

GMM Gaussian Mixture Model

GSS General Smoothing Splines

G2P Grapheme-to-Phoneme

GVV Glottal Volume Velocity

H1 Amplitude of first harmonic

H2 Amplitude of second harmonic

HMM Hidden Markov Model

HTS HMM based Text to Speech System

HNR Harmonic to Noise Ratio

HTK HMM-based Toolkit

IVRS Interactive Voice Response System

IT Information Technology

IITG Indian Institute of Technology, Guwahati

IITG-HTS HTS for standard Assamese built at IITG

IPA International Phonetic Alphabet

JHU John Hopkins University

KLD Kull-Back Leibler Divergence

LLT Local Linear Transformation

LMR Linear Multivariate Regression

LP Linear Prediction

LPC Linear Prediction Coefficients

LSF Line Spectral Frequency

LSP Line Spectrum Pairs

L2S Letter to Sound

MA Moving Average

MCD Mel Cepstral Distortion

MCEP Mel Cepstral Coefficients

MDL Minimum Distance Length

MFC Mel Frequency Cepstrum

MFCC Mel Frequency Cepstral Coefficients

MGC Mel Generalized Coefficients

MIT Massachusetts Institute of Technology

ML Maximum Likelihood

MLSA Mel Log Spectral Approximation

MOS Mean Opinion Score

MRI Magnetic Radio Imaging

MUSHRA Multiple Stimuli with Hidden Reference and Anchor

NAL Nalbaria

NTD Northern Thai Dialect

OBR Onset of Burst Release

OQ Open Quotient

OV Onset of Voicing

PARCOR Partial Correlation

PID Pitch Intensity Duration

PSD Power Spectral Density

PSNR Peak Signal to Noise Ratio

RAPT Robust Algorithm for Pitch Tracking

RCILTS Resource Centre for Indian Language Technology Solutions

RMSE Root Mean Square Error

SDS Spoken Dialogue Systems

SPTK Speech Processing Toolkit

SPSS Statistical Parametric Speech Synthesis

SSANOVA Smoothing Spline Analysis of Variance

Stem-ML Soft Template Mark-up Language

SWIPE Sawtooth Waveform Inspired Pitch Estimator

S2S Speech-to-Speech

TD Temporal Decomposition

TP-FT Text Prompts for the Formant Transformation module

TP-TTS Text Prompts for the TTS

TP-VC Text Prompts for the Voice Conversion module

TTS Text to Speech

TU Tezpur University

TU-TTS TTS for standard Assamese built at TU

UCLA University of California Los Angeles

UV Unvoiced

V Voiced

VoQ Voice Quality

VQ Vector Quantization

VS Voice Source

VFS Vowel Formant Space

VFS-AIR Vowel Formant Space for AIR

VFS-NAL Vowel Formant Space for NAL

VFS-TTS Vowel Formant Space for TTS

VOT Voice Onset Time

VT Voice Transformation

WAE White American English